UiO : University of Oslo

Maoxin Zhang

# Process data analysis in problem-solving tasks

## Thesis submitted for the degree of Ph.D.

Centre for Educational Measurement
Faculty of Educational Sciences

UNIVERSITAS OSLOENSIS · MDCCCXI ·

2023

*Series of dissertations submitted to the*
*Faculty of Educational Sciences, University of Oslo*
No. 369

ISSN 1501-8962

Cover: UiO.
Print production: Graphic center, University of Oslo.

# Acknowledgements

On August 17th, 2019, I traveled to Oslo from my hometown Chongqing, China, which is 7,500 kilometers away. The distance is even greater than the Earth's equatorial radius of 6,378 kilometers. The reason that I came to Norway is to pursue my Ph.D. The moment I landed in Oslo, I asked myself what would happen during the four years and whether I could survive here. Now, four years later, I am delighted to share my answers to these questions. It has been an enjoyable time at the University of Oslo, and these four years have been an enriching period of my life. I have had great opportunities to do my research under excellent supervision, met congenial colleagues and friends, and also enjoyed the natural beauty of Norway. Along the journey, there are so many people who have supported me and contributed to the completion of this dissertation.

First and foremost, I am deeply grateful to my esteemed supervisors (the best supervisors in the world), Dr. Björn Andersson and Dr. Samuel Greiff. Their extensive knowledge, brilliant ideas, and constructive feedback have significantly improved my dissertation. Their patience, support, and encouragement throughout my Ph.D. journey are greatly appreciated. Björn, I would like to express my sincere gratitude for your dedicated supervision over the past four years. The 120 supervision meetings reflected in my calendar have been instrumental in shaping my research ideas, clarifying statistical questions, and addressing reviewers' comments. These interactions have epitomized an ideal supervisor. My gratitude also goes to Samuel, whose expertise in problem-solving has broadened my understanding of the field and significantly improved the literature review of the dissertation and Article II. Even though you are busy with many responsibilities, you never hesitate to provide invaluable guidance in addressing the critical comments from the reviewers and encourage me to keep going.

Second, I would like to thank the reviewers who have provided thoughtful comments on our papers, some of which were truly painful. Despite the discouraging rejections (Article I was rejected three times and Article II four times), the encouragement of my supervisors and friends, and the understanding that criticism is meant to improve, kept me going. To those who are currently suffering from paper rejections, you are not alone, and your work will indeed find a home.

Third, I would like to express my gratitude to my friends and colleagues. I owe special thanks to Xin Guan and Jiaying Xiao, who are also pursuing their Ph.D.s in Japan and the U.S., respectively. Since our time together at Chongqing No. 1 Middle School in 2009 and Beijing Normal University in 2012, your continued support has meant a lot to me. The geographical distance and time differences have not hindered our daily exchanges. I am grateful to have friends in Oslo as well, Jing Wei, Junyi Yang, Wen Zhong, and Qi Qin, whose company has helped reduce my homesickness. I also owe a debt of gratitude to Ahmet Acer, who celebrates life's joys with me and offers comfort in its challenges. In addition, my journey has been enriched by the support and company of my dear

CEMO colleagues, including Sigrid Blömeke (R.I.P.), Rolf Vegar Olsen, Ronny Scherer, Stefan Schauber, Johan Braeken, Tara Sarin, Amanda Marie Grini, Malene Jordal Aase, Grunnhild Nedberg Grønlid, Chia-Wen Chen, Wangqiong Ye, Saskia van Laar, Jose Manuel Arencibia Aleman, Diego Gonzalez Campos, Fredrik Helland-Riise, Jarl Kleppe Kristensen, Yuriko Kirilovna Sosa Paredes, Henrik Galligani Ræder, Jelena Veletic, Denise Reis Costa, Kseniia Marcq, Tony Tan, and Haakon Eidem Haakstad.

Last, I extend my heartfelt appreciation to my family for their emotional and unconditional support, and to myself for insisting on completing this journey. Looking back on the decision to pursue my Ph.D. at the Centre for Educational Measurement, University of Oslo, Norway, I am confident that future reflections will be filled with a sense of accomplishment, fulfillment, and well-being.

**Maoxin Zhang**
Oslo, October 2023

# Abstract

Problem-solving skills, especially problem-solving skills in technology-rich environments, are critical in today's world as one of the 21st-century skills. International large-scale assessments, such as the OECD's Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC), have highlighted the importance of problem-solving and included it as a core domain in their assessments. These assessments have been widely administered on computers, producing performance data and process data that record a detailed history of the human-computer interaction. Examples of process data in our studies consist of action sequences and timestamps for each action (e.g., mouse clicks and keystrokes).

This doctoral dissertation aims to deepen the understanding of problem-solving through the analysis of process data that capture problem-solvers' response processes. The thesis begins with an extended abstract. The extended abstract offers a comprehensive overview of the project and delves into a literature review, covering topics such as problem-solving theories, process data analysis, and latent variable model estimation. Furthermore, this abstract presents the theoretical foundations and methods used, summarizes the main findings, and concludes with a discussion of the contributions and limitations of the project.

The second part of this dissertation consists of four articles. To improve the understanding of problem-solving, we analyze process data from PISA 2012 and PIAAC 2012 with the aim of identifying solution patterns (Article I) and validating cognitive processes involved in problem-solving (Article II) within the framework of latent variable modeling. Both Articles I and II define different types of process-based measures, and analyzing them together increases the computational burden. To provide a fast estimation method for high-dimensional latent variable models, we propose to use higher-order Laplace approximations in Articles III and IV. In addition, our approach can simultaneously account for a mixture of ordinal, continuous, and count variables, as well as the dependencies of observed variables from the same item.

In summary, this dissertation highlights the potential of process data to improve the understanding of how respondents solve problems and provides tools to increase the estimation efficiency when modeling process data and performance data jointly within the framework of generalized linear latent variable models. Our findings can potentially benefit educational practice by helping students reflect on their response processes, aiding teachers in tailoring their instructions for different students and tasks, and providing materials for test developers to validate task design and for training programs aiming at improving problem-solving skills.

# List of Papers

## Paper I

Zhang, M.*, & Andersson, B. (2023). Identifying problem-solving solution patterns using network analysis of operation sequences and response times. *Educational Assessment*, *28*, 172-189.

## Paper II

Zhang, M.*, Andersson B., & Greiff S. (2023) Investigating planning and non-targeted exploration in PIAAC 2012: Validating their measures based on process data and investigating their relationships with problem-solving competency. *Journal of Intelligence*, *11*, 156.

## Paper III

Andersson, B.*, Jin, S., & Zhang, M. (2023). Fast estimation of multiple group generalized linear latent variable models for categorical observed variables. *Computational Statistics & Data Analysis*, *182*, 107710.

## Paper IV

Zhang, M., Andersson, B.*, & Jin, S. Estimation of generalized linear latent variable models for performance and process data with ordinal, continuous, and count observed variables. *Submitted to British Journal of Mathematical and Statistical Psychology.*

# Contents

# List of Figures

# List of Tables

Chapter 1

# Introduction

## 1.1 Background

People have been continually faced with problems in both personal and professional contexts. For example, students need to find the answers to exam questions and workers need to complete professional tasks such as designing a product for customers. That is, people are often required to engage in problem-solving activities. As a result, problem-solving skills have been identified as one of the 21st-century skills that are highly relevant to success in today's world. Studies have demonstrated that problem-solving skills or competency are highly relevant to school achievement (Veerasamy et al., 2019), job performance (Autor et al., 2003), and well-being (Aburezeq & Kasik, 2021). In the information age, problem-solving skills in a digital world are becoming especially important. In today's world, information is widespread and rapidly exchanged, and digital technologies greatly facilitate people's lives. This demands people acquire knowledge about how to handle specific digital tools.

As technology has evolved, the assessment of problem-solving has experienced a shift from self-report questionnaires to interactive, digital assessments. Traditional assessments of problem-solving have applied, for example, questionnaires such as the Problem-Solving Style Inventory including 24 items on six dimensions (Cassidy & Long, 1996), analytic problem-solving tests including a description of problems and multiple-choice or open-ended questions (OECD, 2003), and think-aloud protocols to measure the problem-solving processes (Wolcott & Lobczowski, 2021). The development of technology has accelerated computer-based assessments that display tasks on computer screens and allow respondents to interact with the computer (Greiff et al., 2013). Computer-based problems provide a digital environment for respondents to interact with the stimulus, such as clicking on buttons with a mouse and typing text on a keyboard. All the operations performed by individuals on the computer can be recorded in log files along with timestamps. In effect, log files can depict the history of human-computer interactions and allow researchers and educators to "read" respondents' problem-solving processes (i.e., how respondents approach a problem step by step), thus playing an increasingly important role in measuring people's problem-solving skills. However, it is challenging to analyze log files due to their complex structure and the integration of information from action sequences and response times.

This Ph.D. project focuses on large-scale, computer-based assessments of problem-solving competency and makes use of the information recorded by computers, namely, log files. By mining the information embedded in log files, this project aims to better understand how respondents solve problems and to make methodological contributions to joint models of information from different aspects of log files (e.g., response times, actions) and performance data.

## 1.2 Overarching aim

The overall goal of the project is to gain a deeper understanding of problem-solving through the analysis of process data. The overarching aim encompasses two main elements: understanding problem-solving and using process data. Therefore, the project adopts two different perspectives. First, from a substantive perspective, it seeks to understand problem-solving in terms of the cognitive processes and solution strategies involved in problem-solving. Second, from a methodological perspective, the project focuses on overcoming the challenges of analyzing process data. This includes developing and refining methods for structuring process data, extracting valuable features, and inferring the latent constructs underlying these extracted features. Sound and appropriate methods lay the foundation for reliable findings related to content understanding.

The relationships between the two perspectives and the articles in this dissertation are illustrated in Figure 1.1. The substantive perspective revolves around enhancing our understanding of how individuals solve problems by mining information from log files. Specifically, we aim to identify test-takers' solution patterns for a given task (Article I) and to infer cognitive processes based on features extracted from log files across multiple tasks (Article II).

On the other hand, the methodological perspective aims to overcome the challenges associated with the analysis of process data. When synthesizing various features such as correctness, response times, and action sequences, it often proves challenging to incorporate them into a single model. The difficulties arise for several reasons: a) certain variables violate specific model assumptions (e.g., the discrete variables violate the normal distribution assumption), b) process data exhibit complex dependencies, and c) the high dimensionality increases the computational burden. To address these issues, Articles III and IV strive to bridge this gap by applying computationally efficient algorithms to estimate joint models of the features extracted from log files and performance data.

## 1.3 Overview of the Articles

This dissertation consists of four articles. An overview of these articles is presented as follows.

*Article I: Solution Patterns.* Article I introduces an approach for identifying solution patterns using detailed action sequences and response times. We borrow techniques from social network analysis to visualize the process data and extract valuable features from network graphs according to the problem-solving theory (Mayer & Wittrock, 2006), cluster the problem-solvers based on the extracted features using latent profile analysis (Gaussian Mixture Models), and illustrate the proposed approach using a problem-solving task from the 2012 version of the Programme for International Student Assessment (PISA). Our approach has the

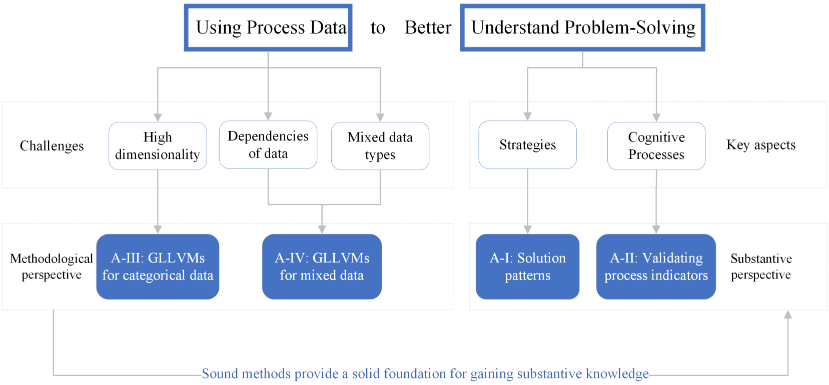

Using Process Data
to
Better
Understand Problem-Solving
Challenges
High dimensionality
Dependencies of data
Mixed data types
Strategies
Cognitive Processes
Key aspects
Methodological perspective
A-III: GLLVMs for categorical data
A-IV: GLLVMs for mixed data
A-I: Solution patterns
A-II: Validating process indicators
Substantive perspective
Sound methods provide a solid foundation for gaining substantive knowledge

Figure 1.1: The overarching aim of the dissertation.

potential to benefit educational practice. Specifically, students can review and reflect on their individual problem-solving processes, teachers can adapt their instructions for students with different solution patterns, and test developers can validate their task design by comparing the solution patterns with their expected ones.

*Article II: Validating process indicators.* Article II focuses on two important cognitive processes involved in problem-solving: planning and non-targeted exploration. Planning refers to mental simulations of future activities, and non-targeted exploration refers to exploratory behaviors that seek information that is not necessary to solve the problem. We examine the internal construct validity of the process indicators for planning (based on response times) and non-targeted exploration (based on actions) using seven problem-solving tasks from the 2012 version of the Programme for the International Assessment of Adult Competencies (PIAAC). In addition, we estimate the overall and task-specific relationships between planning, non-targeted exploration, and problem-solving competency. Confirmatory factor analysis is applied to analyze the categorized process indicators. Our findings provide evidence for the validity of the process indicators and offer insights into the functions of planning and non-targeted exploration in dynamic, information problems.

*Article III: GLLVMs for Categorical Data.* Article III develops a computationally efficient estimation method for generalized linear latent variable models (GLLVMs) for categorical data. The proposed estimation method applies a second-order Laplace approximation to the marginal likelihood estimation, which can greatly increase the estimation efficiency compared to quadrature-based methods and recover model parameters better than a first-order Laplace approximation, according to the results of our simulation study. This approach can deal with high-dimensional models, complex model structures including cross-loadings, and multiple groups.

*Article IV: GLLVMs for Mixed Data.* Article IV proposes a computationally efficient estimation method to jointly model ordinal data, continuous data, and count data within the framework of GLLVMs. We apply first- and second-order Laplace approximations to efficiently approximate the integrals of the marginal likelihood function and investigate their performance through simulation studies. The proposed approach can be applied to the joint analysis of performance data (ordinal responses) and process data (response times and the number of actions) from computer-based assessments and other measurement tools resulting in a combination of different data types.

It can be seen that Articles I and II shed light on the problem-solving processes and answer the question of how problem-solvers solve problems, but the scope of the view has been broadened: a) from analyzing a single task (Article I) to generalizing across multiple tasks (Article II), and b) from describing the observed behavioral patterns (Article I) to inferring the unobserved cognitive processes (Article II). Articles I and II both extract process indicators from log files, which capture the essential information about problem-solving processes. These indicators describe respondents' interactions with the computer from different aspects, but it is difficult to analyze them simultaneously and efficiently in a single model due to high dimensionality and different data types. Articles III and IV then aim to overcome these challenges by using Laplace approximations. Specifically, Article III focuses on categorical data and provides a basic procedure for applying higher-order Laplace approximations to the estimation of GLLVMs. Article IV then applies the procedure described in Article III and extends the type of indicators to a mixture of ordinal, count, and continuous variables.

## 1.4 Outline of the dissertation

The dissertation is divided into two parts: an extended abstract and four articles. The extended abstract aims to give an overview of the dissertation, provide theoretical and methodological reflections on the articles, and discuss the contributions and limitations of the work.

The extended abstract consists of six chapters and is organized as follows. Chapter 1 (i.e., the current chapter) provides a general overview of the research topics, the overarching aims of the dissertation, and a brief summary of the articles. Chapter 2 reviews the relevant conceptual and empirical literature on problem-solving, process data analysis, and latent variable model estimation. Specifically, I briefly summarize the literature on problem-solving theories from a broad perspective, including the definition of problems and problem-solving, research on cognitive processes and problem-solving strategies, and the assessment of problem-solving competency. Similarly, the definition, challenges, and analysis methods of process data are reviewed. Since the articles in the dissertation all use latent variable models, I also present the existing approaches for estimating latent variable models. Chapter 3 explains the theoretical foundations for the empirical studies in Articles I and II, including the problem-solving theory that we directly employ in the articles and the rationale for using process data to help infer unobserved mental processes. It also describes how Articles I and II are developed from the theoretical foundations. Chapter 4 outlines the methods used in the dissertation. Specifically, it describes the data sources, the measures that we define from the process data, and the modeling of the measures. Some ethical considerations are also presented. Chapter 5 summarizes the articles. Chapter 6 discusses the theoretical, practical, and methodological contributions of the dissertation in addition to the limitations.

Chapter 2

# Literature Review

In this chapter, I introduce and summarize literature related to the dissertation in terms of research on problem-solving, process data analysis, and latent variable models.

## 2.1 Research on problem-solving

### 2.1.1 Definition of problems

The word "problem" is derived from a Greek word meaning obstacle (Jonassen, 2010). The obstacle lies between the current state (what people know) and a desired state (what people want to achieve). This defines the first attribute of problems. Another attribute refers to the social, cultural, or intellectual worth embedded in achieving the goal (Jonassen, 2000). That is, the problem should be worth solving for the problem-solvers. Problems can be, for example, how to get good grades on school exams, how to plan a holiday trip, and how to increase work efficiency.

### 2.1.2 Nature of problems

Problems vary in structuredness, dynamicity, domain specificity, routine, context, and complexity (Jonassen, 2010). Below, I briefly describe the nature of a problem from these aspects and give examples of problems that fall into each category.

Structuredness describes the clarity of a problem (Arlin, 1989). A well-structured problem presents all the needed information to problem-solvers including a well-defined initial state, the rules of operators, and a known desired state (Wood, 1983). For example, jigsaw puzzles are well-structured problems. An ill-structured problem, on the other hand, lacks one or more elements of the needed information. For example, the problem of designing a house lacks information about the desired state, and the operator rules are unclear.

Dynamics refers to how problems are displayed in the system. Problems can be displayed in a static system or in a dynamic system. In static problems, all information that is necessary to solve the problem is present at the outset and will not change over time (Jonassen, 2000). In contrast, the information is gradually revealed or changes over time in a dynamic problem (Stadler, Niepel, et al., 2019). The problem situation of dynamic problems can be changed in various ways, such as through the interventions of problem-solvers and the eigendynamics of the system itself. An eigendynamic change is the automatic increasing accrual of interest on a bank account. Dynamics is a key feature of complex problem-solving (Stadler, Niepel, et al., 2019).

Domain specificity indicates if the problem requires domain-related knowledge. Some problems rely on specific strategies within a domain, such as physics and biology, and are called domain-specific problems. The assignments in a specific textbook are usually domain-specific problems. The other type is domain-general problems that require a cross-curricular skill in and of itself (Greiff et al., 2014). For example, making a trip plan within a limited budget is a domain-general problem, which needs travelers to have basic financial management knowledge, information searching skills, and communication skills.

Routine and non-routine problems are distinguished by whether the problem solver has developed a ready-made solution procedure (Mayer, 1998). That is, whether the problem is routine or not depends on the problem solver's familiarity with the specific problem. For instance, a calculus problem for graduates majoring in mathematics is routine, but it is non-routine for elementary students. If we consider the definition of a problem (Jonassen, 2000), the solution to a routine problem is not vague; therefore, a routine problem is more like a task than a problem. If one intends to solve a non-routine problem, the individual would need to invent a novel way of approaching the problem. Therefore, non-routine problems are also called creative problems (Mayer, 1999).

The context of a problem refers to the situatedness described in the problem (Rehm et al., 2003). Namely, the situation in which the problem occurs. This can include everyday problems, such as buying a drink from a vending machine, or workplace problems, such as engineers trying to figure out why a factory machine is malfunctioning.

It is worth noting that these characteristics of problems are all more or less on a continuum rather than an absolute binary classification. For example, a problem's structuredness indicates the extent to which the problem is well clarified, and a problem can contain both domain-specific and domain-general components. For example, solving math problems requires domain-specific skills, such as knowledge of algebra and geometry, and domain-general skills, such as breaking a problem into subproblems and reflecting on the solution.

### 2.1.3 Problem-solving theories

In everyday life, at school, and at work, different types of problems always arise, requiring people to develop strategies to solve them. The cognitive process of transferring a given state into a goal state when the solution is opaque is called *problem-solving* (Mayer & Wittrock, 2006).

Historically, the research related to problem-solving in education and psychology is primarily rooted in three basic theoretical approaches (Mayer, 1999, 2019): associationism (Mandler & Mandler, 1964), Gestalt psychology (Wertheimer & Wertheimer, 1959), and information processing theory (Simon & Newell, 1971).

The associationism approach emphasizes the associations between the elements of the cognitive representations, and the associations can be strengthened or weakened by positive or negative reinforcement (Thorndike, 1911). From this point of view, problem-solving is essentially to apply trial-and-error until accidental success (Mayer, 2019), and the operations that lead to successful solutions are reinforced. For example, to solve a word puzzle, such as solving the anagram problem of converting "rdow" to "word", the associationists will continually attempt to change the order of the given anagram until reaching a known common word. In the associationism approach, problem-solving does not involve creative thinking activities.

Different from the associationism approach, the Gestalt psychology approach values reproductive thinking and gaining insight into the problem (Wertheimer & Wertheimer, 1959). Reproductive thinking involves seeking a familiar problem that is similar to the current problem and transferring the solution to the present situation. Another key concept, insight, refers to a deep understanding of a problem and often comes along with the Aha! experience - a sudden and obvious revelation (Danek et al., 2016). Gestalt psychologists typically focus on ill-structured problems that require respondents to creatively restructure their cognitive representations of the problem. An example problem is to use six equal-length sticks to construct four equilateral triangles (Mayer, 1999). Insight into this problem can be gained when a respondent suddenly realizes that the problem can be solved by constructing a pyramid in a three-dimensional space.

An extensive analysis of problem-solving begins with cognitive psychology from an information-processing approach (Simon & Newell, 1971; Wood, 1983). The information-process approach assumes that we can use a machine (e.g., computer) to simulate the cognitive processes of humans (Simon, 1979). The information-processing theory assumes that information is processed serially, namely one process at a time, and that the input and output of the processes are stored in short-term memory with a limited number of symbols (Simon & Newell, 1971). However, it is also possible to retrieve information from long-term memory. Examples of information-processing models are the Logic Theorist (Newell & Simon, 1956) and the General Problem Solver (Newell et al., 1959) programs. In these programs, the objects and operators are well-defined and the key to a program is to find the sequence of operators transforming the initial state into the goal state. The task of a problem-solver is then to discover and understand the information of the problem environment. To do so, problem-solvers need to represent the problem environment in their mind as a space (called *problem space*) of various possible states of the problem (Simon & Newell, 1971). There are many nodes in a problem space, each representing a particular state of knowledge - that is, what the problem-solver understands about the problem at a specific moment (Simon & Newell, 1971). Problem-solvers then search for a solution throughout the problem space until they reach a knowledge state that covers the problem solution. Therefore, diving into the structure of problem spaces is crucial according to the information-processing approach. Further development of the problem space is to distinguish a rule-space and an instant-space (Simon & Lea, 1974). A rule-space refers to possible rules of the problem, and an instant-space consists of possible states of the problem.

Based on the fundamental theories, researchers have comprehensively explored problem-solving from diverse perspectives. Studies from an educational standpoint have focused on testing instructional strategies to enhance students' problem-solving skills (Gallagher et al., 1992). Suryanto et al. (2021) investigated the impact of social skills on problem-solving, from a perspective of social interaction. On the neuroscience front, researchers have examined the brain activities during problem-solving (Unterrainer et al., 2003). Additionally, from an artificial intelligence perspective, attempts have been made to simulate the human problem-solving process (Ouyang et al., 2023). This project aims to gain insight into the cognitive processes when solving a problem by using the available information in process data. In the following subsection, I summarize relevant problem-solving models that describe the cognitive processes involved in problem-solving.

### 2.1.4 Cognitive processes in problem-solving

Researchers have proposed several models to describe cognitive problem-solving processes. Below, I introduce problem-solving models from the approaches of Gestalt psychology, information-processing, complex problem-solving, and mathematical problem-solving. Based on the existing theories, some international large-scale assessments adapt their theoretical framework of cognitive processes by combining certain processes and adjusting the focus of the cognitive processes according to their purpose. This subsection also presents the theoretical frameworks of the problem-solving domain in PISA and PIAAC.

#### 2.1.4.1 Gestalt psychology perspective

Gestalt psychologists emphasize the importance of analytic thinking in problem-solving and outline four stages of the problem-solving process (Weisberg, 2015): a) finding a solution through transfer by comparing prior knowledge and experience in other similar problems and applying the same solution to the current problem, b) finding a solution through heuristic methods such as trial-and-error and breaking down the problem into more accessible subproblems (Mayer, 2019), c) finding a solution through restructuring by incorporating new information found in the problem and re-analysis the problem, and d) finding a solution through insight by creatively reconsidering a novel type of solutions. The cognitive processes proposed by Gestalt psychologists are more relevant to solving non-routine or creative problems. The Gestalt model is criticized because it is based on an introspective method and lacks reliability and validity (Schoenfeld, 2016).

#### 2.1.4.2 Information-processing perspective

The general problem-solving process in information-processing models specifies a) the understanding process and b) the searching process (Simon & Newell, 1971). The understanding process refers to the process in which respondents attempt to understand a problem statement before trying to solve it (Simon, 1979). The searching process refers to selectively searching through the problem space. The information-processing theory assumes that respondents search sequentially and gradually add successive accretions to the problem space (Simon & Newell, 1971).

Experts outperform novices in recognizing the problem space and require less time to complete the searching process (Jonassen, 2010).

In addition to the two cognitive processes (Simon & Newell, 1971), the information-processing approach also developed other models of problem-solving processes. For example, the IDEAL (identify, define, explore, act, look) Problem Solver model (Bransford & Stein, 1984) proposed five problem-solving processes including identifying the potential problems, defining the problems, exploring potential strategies, acting on the strategies, and looking back to evaluate the effects of the actions.

The information-processing models are examined using information-processing programs such as the General Problem Solver and the Logic Theorist programs, which differ from the introspective method in the Gestalt psychology approach. However, these information-processing models have weaknesses in three aspects. First, information-processing models are less suitable for ill-structured problems because the goal state of the problem and the potential operators are not entirely clear in ill-structured problems. Second, information-processing models tend to propose a uniform procedure of problem-solving, but problems vary in domain, content, and form, which is not accounted for in the general models. Last, the models lack specific and explicit suggestions on how to solve problems (Jonassen, 2000).

#### 2.1.4.3 Complex problem-solving perspective

The information-processing models are more suitable for well-structured, static, and academic problems, whereas the emphasis on problem-solving has shifted to more ill-structured, dynamic, and everyday problems since the 1970s (Wenke et al., 2005). Some theoretical concepts of the information-processing approach, such as the problem space and internal representations of problems, have been expanded and applied to complex problem-solving since the twenty-first century (Fischer et al., 2011). In complex problem-solving, problem-solvers need to systematically interact with the problem to acquire information about the problem, because certain pieces of information required to address the problem may be missing or incomplete (Funke, 2001).

The cognitive processes of complex problem-solving generally contain two phases: knowledge acquisition and knowledge application (Funke, 2001; Greiff et al., 2013). Knowledge acquisition, also known as system identification, refers to finding out the details of the problem environment, such as the functions of available buttons on the screen (similar to an instant-space in information-processing) and the connections between the variables (called structural knowledge) in the system that is inferred from the instant-space (Funke, 2001). Namely, this process is exploratory and aims to gain knowledge about the problem environment as well as the relationships among the elements in the environment or system.

After figuring out how the machine works, knowledge application, also known as system control, then describes applying the acquired knowledge to solving the current problem (Funke, 2001). The knowledge application process requires respondents to utilize internal representations of the problem and monitor the problem-solving progress (Fischer et al., 2011). More specifically, in this process, problem-solvers often need to predict the dynamics of the problem system according to prior knowledge and the knowledge acquired in the previous process. In addition, problem-solvers need to track the progress and consider the feedback from the problem system. Sometimes, problem-solvers may reach an impasse (Ohlsson, 1992) or perceive that the progress rate is too slow (MacGregor et al., 2001), and they may change or restructure their mental representation of the problem. That is, it is possible for problem-solvers to switch back to the knowledge acquisition process.

#### 2.1.4.4 Mathematical problem-solving perspective

In addition to general problem-solving models and complex problem-solving models, researchers have also delved into domain-specific problem-solving. For example, Polya (2004) particularly focused on mathematical problems and proposed the following four phases to solve mathematical problems. Begin by understanding the problem and its nature. In this step, respondents identify the initial state of the problem by asking, "What elements are present?" and the goal state by asking, "What do I want to achieve?" They then synthesize the information gathered into a mental representation and devise a strategy. This may involve applying a solution from a previously known problem or creating a new insight (Mayer, 1999). Next, they carry out the formulated plan. Finally, respondents reflect on and evaluate the rationale for the solution. In addition to the theory of cognitive processes in mathematical problem-solving, Polya (2004) also offers specific strategies for solving problems, some of which are introduced in the next subsection. These strategies have been implemented in mathematics education.

#### 2.1.4.5 PISA and PIAAC framework

Integrating the existing theories (e.g., Greiff et al., 2014; Mayer & Wittrock, 2006; Polya, 2004; Simon & Newell, 1971; Wüstenberg et al., 2012), PISA and PIAAC have adapted their theoretical frameworks of cognitive processes in problem-solving according to their focus, and these frameworks also guide the task design aiming at measuring problem-solving competency. Their frameworks synthesize theories and empirical studies in problem-solving research, especially research on complex problem-solving (Fischer et al., 2011), and combine specific cognitive processes described in previous literature to serve as a measurement framework (OECD, 2019). Problem-solving in PISA and PIAAC has coherent associations and is similar in their theoretical frameworks of the problem-solving domain. Here, I introduce the theoretical frameworks of problem-solving provided by PISA (2003, 2012, and 2015) and PIAAC (2012, 2022).

*PISA 2003.* The items from the PISA 2003 problem-solving domain are based on personal life, work and leisure, and community and society contexts and involve a wide range of disciplines (OECD, 2003). The assessment includes relatively well-structured problems delivered in the form of pencil-and-paper format, which are static problems. Six problem-solving processes were proposed in the PISA 2003 problem-solving domain: understanding the problem, characterizing the problem, representing the problem, solving the problem, reflecting on the solution, and communicating the problem solution (OECD, 2003). It is worth noting that the processes may occur in different orders and may not occur for certain problems (OECD, 2003).

*PISA 2012.* Creative problem-solving is a domain in PISA 2012. The problems in PISA 2012 are based on a personal or social context and are mainly domain-general problems. The assessment includes both static (30%) and dynamic (70%) problems, as well as well-structured and ill-structured problems. The theoretical framework for creative problem-solving comprises four cognitive processes. First, explore and understand the problem by observing and interacting with the problem situation, searching for information, and identifying obstacles. Second, represent and formulate the problem by creating graphical or verbal representations, developing hypotheses, and organizing information. Third, plan and execute the solution by setting goals, devising a strategy, and implementing the plan. Finally, monitor and reflect on the solution by reviewing intermediate and final results, taking remedial actions, and evaluating assumptions and alternatives (OECD, 2014b). These cognitive processes are not sequential but rather parallel information processes that occur throughout the problem-solving activities of the participants (Lesh & Judith, 2007), which differ from the serial process assumed in the information-processing theory (Simon & Newell, 1971).

*PISA 2015.* PISA 2015 assesses collaborative problem-solving by introducing a computer agent that interacts with participants. All items are dynamic and domain-general, and both well-structured and ill-structured problems are considered. The PISA 2015 framework for collaborative problem-solving consists of four cognitive problem-solving processes in individual problem-solving as in PISA 2012 (OECD, 2014b) and three collaborative problem-solving aspects, including establishing and maintaining shared understanding, taking appropriate actions, and establishing and maintaining team organization (OECD, 2017).

*PIAAC 2012.* The PIAAC 2012 assessment includes the Problem Solving in Technology-Rich Environments (PS-TRE) domain. PS-TRE uses information problems that demand information and communication technology (ICT) skills (OECD, 2019). All items are computer-based and dynamic, involving multiple software applications or pages. The assessment comprises seven ill-structured and seven well-structured problems. The PS-TRE framework includes four cognitive processes: a) setting goals and monitoring process, b) planning and self-organising, c) acquiring and evaluating information, and d) using information. Compared to previous assessment frameworks, PIAAC 2012 emphasizes using information.

*PIAAC 2022.* The second PIAAC cycle shifts the focus from the use of software applications to the adaptability in concurrently solving multiple problems (Greiff et al., 2017). The expert group recommends interactive, dynamic, domain-general, and information-rich problems. The theoretical framework of PIAAC 2022 consists of three major stages within the participants' internal world: problem definition, solution search, and solution application. Every stage entails various cognitive processes (e.g., searching for information and retrieving relevant background information) and meta-cognitive processes (e.g., setting goals and monitoring progress) (Greiff et al., 2017). As participants are expected to dynamically and flexibly adapt their strategies, the cognitive processes involved in adaptive problem-solving are more complicated than those of PIAAC 2012 (OECD, 2021).

As described above, researchers have proposed many theoretical frameworks to describe problem-solving processes. The cognitive processes described by these frameworks share great similarities but use different terminology. In this dissertation, I primarily used the problem-solving theory proposed by Mayer and Wittrock (2006) that describes a general framework of cognitive processes and is widely used in the research on problem-solving now. I will introduce it in more detail in Section 3.1.

### 2.1.5 Problem-solving strategies

In addition to problem-solving processes, researchers have also explored strategies that guide problem-solving. In this subsection, I introduce several widely-used strategies discussed in the research.

No matter which problem-solving theory is adopted, representing a problem is the first step to take. Unlike constructing a general problem space, schema (Gick, 1986) is a knowledge-based representation that originated from Gestalt psychology. A schema is defined as a cluster of knowledge associated with a particular type of problem, including knowledge of the common goals, typical procedures for solving this type of problem, and some constraints (Gick & Holyoak, 1983).

Schemas exist in the problem-solvers' memory system, including domain-specific schemas such as mathematical theorems, and domain-general schemas such as breaking a problem into subproblems that are easier to manage. When a schema is activated during the representing process, a problem-solver can rely on the schema to find the solution without extensive searching activities. For example, to solve a math problem of finding the third side of a right triangle given the lengths of two sides, a schema that includes the Pythagorean Theorem can be activated to guide students through the problem. However, schemas are not always available if a problem-solver does not recognize the particular problem type or has no prior knowledge or experience relevant to the problem type.

The information-processing approach proposes two principal problem-solving strategies: the mean-ends analysis and planning (Newell & Simon, 1956). The means-ends analysis involves a) searching for the difference between the current and desired status, b) identifying and applying operators that can eliminate the differences, and c) if the difference proves to be particularly challenging to eliminate, applying operators, even if they introduce new but more manageable differences (Newell et al., 1959). The means-ends strategy works effectively in well-structured problems but might be infeasible in ill-structured problems. The second strategy - planning - involves a) abstracting the initial status and operators by omitting some details, b) trying to solve the abstracted subproblems, and c) conceiving a plan for the original problem based on the successful solution to the abstracted subproblems and executing it (Newell et al., 1959). Note that the planning method may result in no, single, or several plans that may succeed or fail to solve the problem.

Some other heuristic strategies have also been discussed. For example, Polya (2004) introduced heuristics aiding problem-solving in mathematics, such as inference by analogy and discovering general laws from specific examples by induction. Another widely-used heuristic strategy is the generate-and-test strategy (Klahr, 2000), also known as trial-and-error, which means simply applying available operators to the current state and testing whether the problem has been solved. If the desired state has not been achieved, the problem-solver continues applying alternative operators and testing again until reaching the goal. The hill climbing strategy is also heuristic and uses the metaphor of climbing a hill (Simon & Newell, 1971). Suppose you are climbing a hill with different paths, and your goal is to reach the summit. In that case, a helpful strategy would be to always choose the path that leads upwards. In other words, always select the operator that gives the greatest increment.

The varying-one-thing-at-a-time (VOTAT) strategy (Tschirgi, 1980) is often considered an optimal strategy in complex problem-solving tests (Gnaldi et al., 2020; Lotz et al., 2022; Stadler, Fischer, et al., 2019). The VOTAT strategy refers to changing one input variable at a time while holding other input variables constant in order to examine the relationship between the input variable and the outcome. Note that the dynamics of complex problems can reflect autonomous changes in the system over time, namely eigendynamics. The varying-nothing-at-a-time (NOTAT) strategy is optimal for detecting such eigendynamic effects by observing the autonomous changes without any intervention (Lotz et al., 2022).

As a complex cognitive activity, problem-solving involves a variety of cognitive processes and may vary across problem-solvers and types of problems. For example, prior domain knowledge such as schema and experience with similar problems, as well as motivation, can mediate problem-solving (Jonassen, 2010). The characteristics of problems, such as well-structured versus ill-defined problems or static versus dynamic problems, also play an important role in explaining problem-solving. How problem-solvers approach a specific problem and to what extent we can generalize our conclusions about the problem-solving process to other tasks motivated the first two studies of the dissertation.

### 2.1.6 Assessment of problem-solving

Since problem-solving plays a vital role in our life and people differ in problem-solving, it is necessary to develop useful tools to measure individual problem-solving competency. Researchers have attempted to measure problem-solving since the last century. For example, several domain-specific problem-based exams were developed in the 1980s and 1990s. One such assessment is the

OverAll Test (Segers, 1997) in business education, which measures students' ability to retrieve relevant knowledge and understand conditional knowledge, such as identifying when and where to access useful tools to solve the problem. Similarly, the WHAT-IF Test (Swaak & de Jong, 1996) presents conditions, actions, and predictions to measure conceptual knowledge in science education. These tests are designed for problem-based learning in a particular curriculum, namely domain-specific problems.

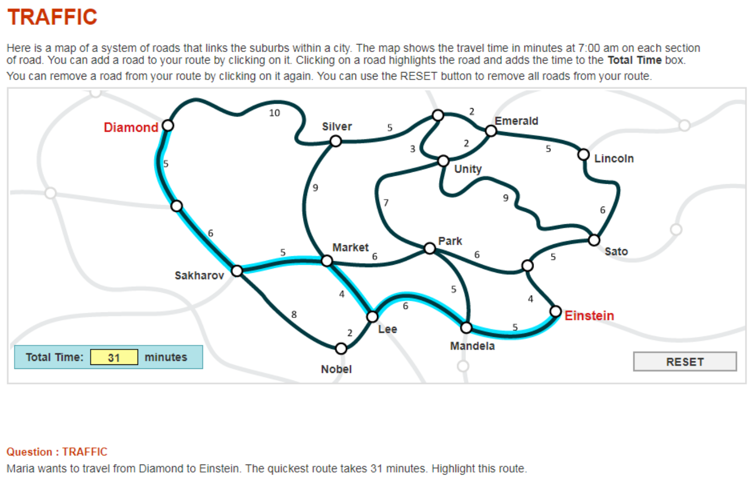
Since the late 1990s, researchers have tended to view problem-solving as an interdisciplinary competency in realistic settings (Mayer & Wittrock, 1996). In line with this thinking, some large-scale assessment projects develop tasks to measure problem-solving competency in an everyday context, such as the Programme for International Student Assessment (PISA) and the Programme for International Assessment of Adult Competencies (PIAAC) of the Organisation for Economic Co-operation and Development (OECD). This section introduces problem-solving assessments in PISA and PIAAC.

#### 2.1.6.1 Problem-solving in PISA

PISA included problem-solving assessments for 650 15-year-old German students in 2000 (Klieme, 2000). The assessment of problem-solving was extended to 41 counties using paper-and-pencil instruments in PISA 2003 where problem-solving is defined as "an individual's capacity to use cognitive processes to confront and resolve real, cross-disciplinary situations where the solution path is not immediately obvious and where the literature domains or curricular areas that might be applicable are not within a single domain of mathematics, science or reading" (p.156; OECD, 2003).

In 2012, PISA began to deliver the items via computers, making it possible for test-takers to interact with computers. Specifically, problem-solving was a focus area in PISA 2012, denoted as creative problem-solving, which is defined as "an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious" (p.30; OECD, 2014b). An example problem of PISA 2012 is presented in Figure 2.1. In this task, the participants were asked to find the quickest route between two locations on the map by clicking on the paths. The computer logs the complete human-computer interaction, providing detailed records of the participants' response processes. Such information is used in Article I to identify the solution patterns of the respondents.

Instead of individual problem-solving, PISA 2015 focused on collaborative problem-solving where students can collaborate with a computer agent. Collaborative problem-solving competency is defined as "the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2017). In PISA 2015, both individual and collaborative problem-solving are assessed.



TRAFFIC

Here is a map of a system of roads that links the suburbs within a city. The map shows the travel time in minutes at 7:00 am on each section of road. You can add a road to your route by clicking on it. Clicking on a road highlights the road and adds the time to the **Total Time** box.
You can remove a road from your route by clicking on it again. You can use the RESET button to remove all roads from your route.

Total Time: 31 minutes

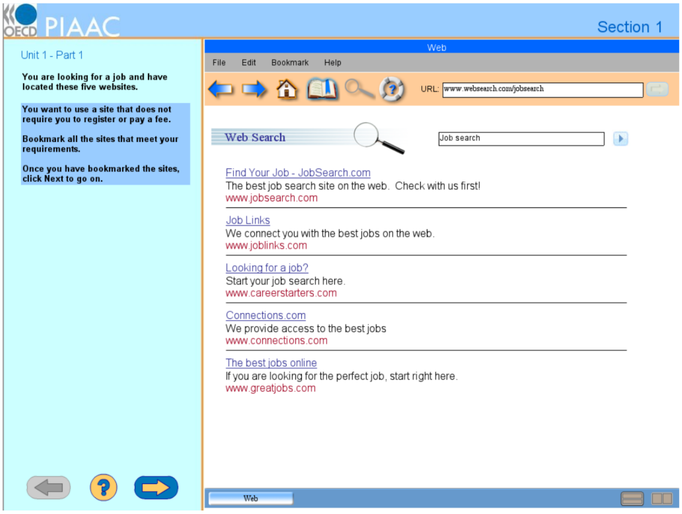
Question : TRAFFIC
Maria wants to travel from Diamond to Einstein. The quickest route takes 31 minutes. Highlight this route.

Figure 2.1: An example problem in the PISA 2012 problem-solving domain. The highlighted route indicates the correct solution.

### 2.1.6.2 Problem-solving in PIAAC

Similar to PISA, PIAAC has also considered problem-solving in the first cycle in 2012 and the second cycle in 2022 and 2023. PIAAC 2012 assesses adults' ability to solve information problems in the domain of problem solving in technology-rich environments (PS-TRE). PS-TRE is defined as "using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (p. 47; OECD, 2012). Similar to PISA 2012, the tasks in the PS-TRE domain were also administered via computers, and the associated log files are publicly available. The PIAAC problem-solving tasks consist of one or more interfaces (Web, email, word processor, and spreadsheet) or pages. An example task of PIAAC 2012 is presented in Figure 2.2. In this task, five Web links are provided, and the respondents are asked to bookmark links that fulfill specific requirements.

After ten years, the second cycle of PIAAC continues the interest in problem-solving, but the focus has shifted from problem-solving in technology-rich environments to adaptive problem-solving (Greiff et al., 2017). Adaptive problem-solving is defined as "a form of problem solving that requires a series of problem reformulations or continual re-evaluation of problem formulations in light of changing conditions" (p.153; Mayer, 2014). An example of adaptive problem-solving is making a trip plan for several family members with different preferences and some constraints like budget (Greiff et al., 2017). The second cycle of PIAAC is still in progress, and the data are not yet available (date until June 2023).



Unit 1 - Part 1

You are looking for a job and have located these five websites.

You want to use a site that does not require you to register or pay a fee.

Bookmark all the sites that meet your requirements.

Once you have bookmarked the sites, click Next to go on.

URL: www.websearch.com/jobsearch

Job search

Find Your Job - JobSearch.com
The best job search site on the web. Check with us first!
www.jobsearch.com

Job Links
We connect you with the best jobs on the web.
www.joblinks.com

Looking for a job?
Start your job search here.
www.careerstarters.com

Connections.com
We provide access to the best jobs
www.connections.com

The best jobs online
If you are looking for the perfect job, start right here.
www.greatjobs.com

Figure 2.2: An example problem in the PIAAC 2012 PS-TRE domain.

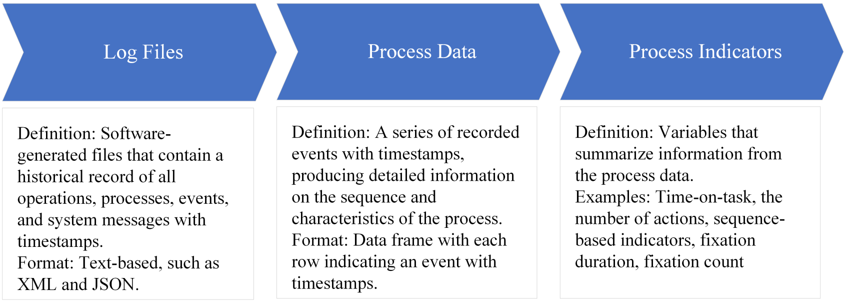
## 2.2 Process data analysis

### 2.2.1 Definition of process data

Many recent assessments have been implemented on computers, which are called *computer-based assessments* (CBA). The item stimulus is displayed on computer screens, and test-takers can interact with the computer by clicking on buttons with a mouse and typing text with a keyboard. Computers can easily generate files that capture the history of everything that test-takers did during the course of the assessment along with timestamps of the operations.

The terminology of such files is mixed in the literature. Some parts of the literature use the terms "click-stream data", "log-file data", or "discrete action protocols" and these terms refer to "log files" in this dissertation. A broad definition of log files is software-generated files that contain a historical record of all operations, processes, events, and system messages with timestamps. An example of log files from a website may include event logs that record users' activities and system logs that document system changes and system errors. In this study, we focus on event logs recording the human-computer interaction and associated timestamps. The format of log files is typically delimited strings of text files, such as Extensible Markup Language (XML) and JavaScript Object Notion (JSON) files.

Text-based files are not directly usable for data analysis, thus requiring converting the text-based strings into a data frame with each row representing a single, time-stamped interaction. The transformed data frame is called *process data* in this dissertation. It is defined as a series of recorded events with timestamps, which provide detailed information about the process of the users'



Log Files

Definition: Software-generated files that contain a historical record of all operations, processes, events, and system messages with timestamps.
Format: Text-based, such as XML and JSON.

Process Data

Definition: A series of recorded events with timestamps, producing detailed information on the sequence and characteristics of the process.
Format: Data frame with each row indicating an event with timestamps.

Process Indicators

Definition: Variables that summarize information from the process data.
Examples: Time-on-task, the number of actions, sequence-based indicators, fixation duration, fixation count

Figure 2.3: The definitions of log files, process data, and process indicators in this dissertation.

operations.

From process data, researchers can further extract useful and more abstract indicators or statistics to summarize the response process using a theory-driven or data-driven approach (see Section 2.2.4.1 for more details). These extracted variables are called *process indicators*. Process indicators can be extracted or created based on process data at an individual item level or aggregated level. The definitions of log files, process data, and process indicators are briefly illustrated in Figure 2.3.

### 2.2.2 Challenges of analyzing process data

Process data contain much more information beyond the final task performance (i.e., task scores). However, it is often challenging to analyze and make use of process data. The special characteristics of process data can explain the difficulty, and six challenges of analyzing process data are outlined below (see Figure 2.4).

First, process data typically have a large volume and include a great variety of variables. To be specific, test-takers can actively interact with the computer, producing hundreds of actions that correspond to many categorical variables in single items. Furthermore, process data often contain the timestamps for each operation. In an example from Article I, a respondent performed over 300 operations within a single task. Namely, compared to a single score indicating task performance, process data have a larger volume and contain more details.

A challenge associated with voluminous process data is high dimensionality. High dimensionality can be attributed to multiple resources of log files (e.g., time-related information, operation-related information, and multiple items) and multi-modal data (e.g., eye-tracking data and cognitive response data). For example, Article II considers planning, non-targeted exploration, problem-solving competency, and the residual factors for six items. This yields nine dimensions in the analysis. High dimensionality increases the computational demands.

The third challenge of process data is the varied lengths across test-takers, as problem-solving strategies are mediated by individual differences (Jonassen,

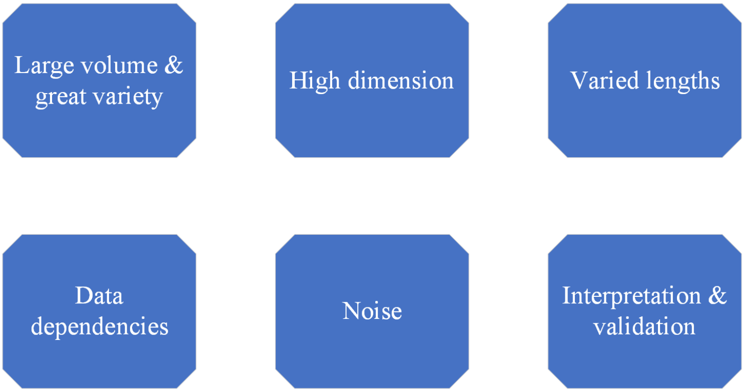


Figure 2.4: The challenges of process data analysis.

2000). Test-takers experience distinct cognitive processes, which are reflected in their interactions with the computer and result in non-fixed lengths of action sequences. For example, the lengths of process data for individual respondents range from 2 (entering and exiting the item directly) to 373 (actively interacting with the computer) on a PISA 2012 item. In contrast, a standardized test often consists of a fixed number of items, and test-takers normally have the same length of response data. Conventional statistical methods, such as factor analysis and item response theory (IRT) models, are less suitable to directly apply to process data.

Fourth, process data describe the ordered sequence of operations, resulting in a complex dependency structure. Specifically, the current operation is related to both the previous and next operations. Moreover, when extracting or creating multiple indicators from the same task, the dependency of the indicators should be taken into consideration as well.

Another characteristic of process data is that they typically contain a great amount of noise (Tang, Wang, He, et al., 2020). For example, test-takers might randomly conduct some actions instead of following a certain strategy. Pre-processing process data and improving data quality are critical for drawing reliable conclusions. A standardized procedure for pre-processing process data, including data cleaning, data re-coding, and missing data handling, has not been established in the literature.

Last, the interpretation and validation of the results of process data analysis can be challenging. For example, the results from a data-driven approach such as machine learning techniques are often difficult to interpret because the analysis is guided by the data rather than by theory. In addition, the extent to which the conclusion can be generalized to other tasks or samples is not well examined, as most studies only focus on a single task and sample.

Despite the above-mentioned challenges of analyzing process data, it is meaningful to overcome the challenges and utilize process data for various purposes. In the next subsection, I describe how process data have been used in empirical studies.

### 2.2.3 Empirical studies of using process data

Process data can provide researchers with valuable information that improves the understanding of respondents' cognitive processes (OECD, 2014b), and the analysis of process data can benefit educational and psychological measurement. Many studies have used process data for various purposes. In general, there are four primary purposes for using process data in the literature: validating test design, identifying response patterns, improving the estimation of the construct of interest, and predicting final responses.

Since process data depict the response process, validating test scores of the assessments is an original usage of process data (Shute et al., 2016; Stoeffler et al., 2020), which can contribute to test development (von Davier et al., 2019). For example, Chung et al. (2002) defined a set of process measures based on behavioral process data and verbal recordings using think-aloud protocols and found that the process measures were significantly correlated with task success. These process measures, such as cause-effect inferences and evaluation of information, are relevant to the dimensions that are expected to be critical to task performance, thus providing evidence of validity for the task.

The second purpose is to reveal response patterns. Specifically, researchers have employed process data to a) distinguish non-effort or disengaged behavior from solution behavior (Y. Liu et al., 2020; Sahin & Colvin, 2020; Ulitzsch et al., 2020), b) identify common problem-solving strategies (e.g., VOTAT and trial-and-error) and profile students (Gao et al., 2022; Gnaldi et al., 2020; Greiff et al., 2018; Stadler, Fischer, et al., 2019), c) compare participants' response process with a pre-defined optimal strategy to examine the extent to which they exhibit a similar action sequence as the expert-defined sequence (He et al., 2021), and d) further examine the relationships between the solution patterns identified and problem-solving competency and task performance (S. Li et al., 2022; Lotz et al., 2022; Vörös et al., 2021) or other personal characteristics, including demographic variables and employment-related variables such as income and work experience (Liao et al., 2019).

Third, researchers have attempted to incorporate process data information into the estimation of the construct of interest. For example, researchers have proposed models to use process data to aid the estimation of problem-solving competency (Chen, 2020; Y. Han et al., 2022; Xiao & Liu, 2023; Zhan & Qiao, 2022). Besides problem-solving competency, researchers have also attempted to infer latent states behind observed actions during the course of problem-solving (Xiao et al., 2021) and to measure other latent traits such as speed (De Boeck & Scalise, 2019).

Fourth, process data have been used in predictive models to predict task performance. That is, researchers extract or create measures/indicators based on process data and use the process indicators as predictors for the success or failure of the task. For example, a study of Chen et al. (2019) predicted the final task score and duration of test-takers based on their event history in process data. Similarly, Z. Han et al. (2019) extracted process indicators and used these indicators to predict the final outcome.

As described in this subsection, process data in problem-solving tasks have gained increasing attention in the field of psychological and educational measurement and have been used for various purposes. However, although process data contains both action sequences and timestamps, the majority of previous studies have centered on either action sequences (e.g., Greiff et al., 2018; He & von Davier, 2016; H. Liu et al., 2018; Ulitzsch et al., 2020) or the time spent on task (e.g., Bolsinova & Tijmstra, 2018; Y. Liu et al., 2020). Recently, more researchers have started employing information from both action sequences and response times to make greater use of log-file information (Chen, 2020; Chen et al., 2019; De Boeck & Scalise, 2019; Ulitzsch et al., 2021; Xu et al., 2020).

This project integrates information from both actions and response times into a single model. The model is used to identify solution patterns (Article I) and reflect on cognitive processes (Article II) from a substantive perspective. Additionally, we develop statistical methods to deal with the challenges of computational burdens when modeling process data and performance data simultaneously (Articles III and IV) from a methodological perspective. As mentioned above, many relevant studies have only focused on a single task (e.g., He & von Davier, 2016; Ulitzsch et al., 2021; Xu et al., 2020; Zhan & Qiao, 2022), and the generalizability of the conclusions is not well examined. This dissertation includes both single-task analyses (Article I) and multiple-task analyses (Articles II to IV).

### 2.2.4 Methods of process data analysis

The methods of process data analysis can be roughly categorized into two procedures: a) extraction of valuable information from process data, and b) application of statistical models using either observed variables only or both observed and latent variables. In the first step, researchers aim to compress a large amount of information contained in process data into a few process indicators, identify the sequential pattern of process data, or visualize process data. The process indicators mirror similar essential information in the original massive data. With the extracted variables, researchers can apply different statistical models for various purposes. In general, statistical modeling of process data focuses on investigating the relationships between the observed indicators, like whether a certain strategy can predict task success, or inferring latent variables from the observed indicators, like whether process data can increase the accuracy of latent ability estimation. In this section, I will briefly summarize the two procedures with examples.

#### 2.2.4.1 Extracting information from process data

*1. Behavioral/time-related process indicators*

Multiple approaches have been applied to extract information from process data. Various behavioral/time-related process indicators, also called features, have been defined in the literature to reflect the response process. These process indicators are defined either through a confirmatory approach or an exploratory approach.

From a confirmatory perspective, feature extraction is based on a certain theory or theoretical framework. An example is from Yuan et al. (2019), where a set of behavioral indices were defined according to the Assessment and Teaching of 21st Century Skills project. Such indicators require extensive effort from experts and are highly task-dependent. Another example is the VOTAT strategy (Tschirgi, 1980) that is described in Section 2.1.5, which has been extensively used in empirical studies (Lotz et al., 2017, 2022; Stadler, Fischer, et al., 2019). This strategy can be identified by inspecting the action sequence and checking if a respondent changes only one variable at a time while keeping all other variables constant to examine the effect of the changed variable. Time-related process measures, such as the first move-latency (i.e., the time interval before conducting the first operation) and the longest duration (i.e., the longest time interval between two successive operations), are proposed to capture the information from response times to reflect the planning process (Albert & Steinberg, 2011; Eichmann et al., 2019). Similarly, some general process indicators at the task level, such as time-on-task (i.e., the total time spent on a task) and the number of actions, are also commonly used in the literature (De Boeck & Scalise, 2019; Gao et al., 2022).

In comparison, feature extraction from an exploratory perspective relies mainly on data rather than theories. Natural language processing (NLP) and dimension-reduction techniques have been adopted in this approach. In NLP, process data are regarded as text-based strings and are analogous to language. He and von Davier (2016) proposed to employ n-grams from NLP to decompose the complete action sequence into smaller units, such as single operations (unigrams) and operation vectors with two (bigrams) or three (trigrams) consecutive operations. Using the action sequence "start, action1, action2, end" as an illustration: unigrams for this sequence are "start", "action1", "action2", and "end"; bigrams are "start, action1", "action1, action2", and "action2, end"; and trigrams are "start, action1, action2" and "action1, action2, end". These grams can be used for subsequent analyses. Furthermore, some dimension-reduction techniques have been used to extract a few latent features from high-dimensional process data. For example, Tang, Wang, Liu, et al. (2020) used a sequence-to-sequence autoencoder technique to transform high-dimensional process data into low-dimensional, numerical latent feature vectors. Similarly, Tang, Wang, He, et al. (2020) applied a multidimensional scaling framework to construct latent features based on dissimilarities between pairwise action sequences. However, the drawback of this approach is that it is rather challenging to interpret the extracted features because no theory is involved in guiding the feature extraction process. In psychology and education, the interpretation and explanation of the results are critical.

*2. Sequence-based methods*

Process data including action sequences and response time are sequential data. Some studies are particularly interested in the sequential relationships and apply sequence mining techniques to extract features to reflect sequential patterns. For example, a sequence mining technique, edit distance, has been used to analyze process data. Edit distance quantifies the distance between two strings by counting the minimum number of operations needed to transfer a given string to the other (Ristad & Yianilos, 1998). These operations can include insertions, deletions, and substitutions. Based on the allowed operations, edit distance can be categorized into, for example, the *Levenshtein distance* that allows deletion, insertion, and substitution and the *longest common subsequence* that allows insertion and deletion (Ristad & Yianilos, 1998). If two strings are similar, only a few operations are needed for the transformation. Namely, edit distance measures the dissimilarity of two strings. Viewing action sequences as strings, edit distance techniques can measure how dissimilar two action sequences are. This technique has been used to quantify the dissimilarity between two action sequences of pairwise test-takers (Tang, Wang, He, et al., 2020) and between an observed action sequence and an ideal action sequence defined by experts (Hao et al., 2015; He et al., 2021). Similarly, the similarities between the time sequences can be computed through the edit distance technique (Ulitzsch et al., 2021). In this way, the action sequence and the time sequence can be converted into distance-based dissimilarity indicators.

Similar to sequence mining, educational process mining methods have also contributed to analyzing sequential data from educational settings. Educational process mining is a branch of data mining, which aims at building a comprehensive process model to reproduce log events (process discovery models), checking behaviors that are deviated from the process model (conformance checking models), and improving the process model (enhancement models) (Bogarín et al., 2018). Among these models in educational process mining, process discovery models are most widely used in practice. Process discovery techniques, such as the fuzzy miner algorithms, construct a model to represent the most likely control flow given the observed activities and produce a graph where vertices represent activities, edges represent the transition between the activities, and edges are weighted by the transition probability (Bogarín et al., 2018). The application of educational process mining centers on online learning activities, but the application to computer-based assessments is limited (Tóth et al., 2017).

Furthermore, the full-path sequence analysis, which is originally used for DNA comparisons, has been applied to analyze process data (Eichmann, Greiff, et al., 2020). In their study, actions were first coded as initial/repeated non-targeted exploration or goal-directed behavior and resetting, and the process data were converted to a sequence of these five categories of actions. The authors then used a full-path sequence analysis and string-matching algorithms on the coded action sequences to cluster participants exhibiting similar behavioral patterns (Eichmann, Greiff, et al., 2020).

*3. Visualization methods*

Another approach has a focus on visualizing process data. Network graphs from social network analysis have been used to visualize the response process based on process data (Vista et al., 2016; Vista et al., 2017; Zhu et al., 2016). A network graph depicts the relationships between vertices. For example, if a class has 30 students and they are asked to nominate their friends in the class, then the network graph includes 30 vertices that represent the students and a number of edges that represent the presence of friendship between two students. The network graph provides a straightforward visualization of friendship within the class. Next, researchers can define network statistics, such as the *density* of the network that describes how closely the students are connected and *centrality* that describes how popular a student is within the class.

Analogous to friendship networks, S. Li et al. (2022), Vista et al. (2016), Vista et al. (2017) and Zhu et al. (2016) viewed individual actions in process data as vertices and the transition between actions as directed edges that point from the previous action to the current action. In addition to a graphical representation of networks, researchers can also define statistics based on networks, called network features, for further inference. For example, two studies measured the importance of each vertex and edge based on *centrality* and frequencies (Vista et al., 2016; Vista et al., 2017). They then identified the prominent action subsequences consisting of important vertices and edges. In another paper by Zhu et al. (2016), the authors defined other network features. To be specific, they weighted the edges with their frequency and computed the density of the network, the importance of the vertex, dyadic local patterns that describe the mutual relationships between two vertices, and sixteen triadic local patterns that describe the relationships among three vertices. Similarly, a recent study employed social network analysis to analyze collaborative problem-solving between two students (S. Li et al., 2022). In their study, they regarded both respondents' operational actions (e.g., mouse clicks) and their chat actions (e.g., ChatA/ChatB indicating that student A/B sends texts in the chat box) as vertices, the transitions of two actions as edges, and the frequency of the transitions as weight. They computed network features using a similar approach as Zhu et al. (2016). These studies showcased the potential of applying social network analysis to visualize response processes and extract network features to reflect the essential information of process data. However, a common limitation is that the time information was ignored in these studies.

Another method is a graph-based data clustering technique (Ulitzsch et al., 2021). The authors first computed the similarities between the action sequences and the time sequences for pairwise participants using edit distance techniques. They then plotted the similarity graph with vertices representing the participants and edges representing their similarity. Finally, the authors performed edge deletion to transform the similarity graph into a simplified graph with a few clusters (Ulitzsch et al., 2021). Respondents within the same cluster were interpreted to exhibit similar behavioral patterns.

#### 2.2.4.2 Statistical modeling of process data

*1. Statistical models using only observed variables*

After extracting or creating process indicators or measures based on process data, researchers can apply statistical models for statistical inference. For this purpose, traditional methods such as regression models and Analysis of Variance (ANOVA) have been used. In regression analysis, process indicators can act as predictors or outcome variables. For example, Liao et al. (2019) used background variables to predict response time, whereas Ren et al. (2019) used behavioral process indicators to predict problem-solving competency and task scores. To compare the features of different groups, the researchers utilized ANOVA to test whether the extracted network features differed across pair compositions (S. Li et al., 2022) and whether the process indicators were different across the clusters identified (Gao et al., 2022). Similarly, with the defined n-grams, researchers applied a weighting method based on frequencies and used a chi-square selection model to detect which grams differentiated successful and unsuccessful test-takers (He & von Davier, 2016; Liao et al., 2019).

In addition, machine learning techniques have been applied for the purposes of regression, classification, and clustering. For example, Z. Han et al. (2019) viewed n-grams as mini-behavioral features and applied random forest algorithms to select the grams that had the highest predictive capability for task success. In a similar vein, Qiao and Jiao (2018) selected features based on their predictive power using four supervised learning techniques including classification and regression trees, gradient boosting, and support vector machine. K-means techniques were applied to group participants based on their process indicators (Gao et al., 2022; Qiao & Jiao, 2018).

*2. Statistical models using both observed and latent variables*

Another approach to statistical modeling is to infer latent variables based on observed process indicators. Existing studies have attempted to incorporate process data into the estimation of problem-solving competency along a continuum or latent classes, infer latent states underlying the problem-solving processes, or jointly model process data and performance data to examine the relationships between problem-solving competency and other latent constructs such as speed.

To incorporate process information into the estimation of problem-solving competency, some traditional psychometric models have been adopted. For example, the modified multilevel mixture IRT model was used to provide ability estimates at a process level and a person level and identify latent classes of the respondents (H. Liu et al., 2018). From a diagnostic perspective, diagnostic classification analysis was applied to process data to estimate problem-solving competency and classify students based on the strategies they used in the task (Zhan & Qiao, 2022). Researchers also applied latent class analysis to profile the respondents based on their process data (Gnaldi et al., 2020; Greiff et al., 2018).

In addition, stochastic process models have been used in process data analysis. A dynamic Bayesian network is a type of stochastic process model, which is a probabilistic graphical model used to model complex systems that evolve over time (Reichenberg, 2018). It assumes that a) the observed activity ($X_t$) at a certain time $t$ is stochastically dependent on latent variables ($\boldsymbol{\theta}_t$), and b) the latent variables ($\boldsymbol{\theta}_t$) are stochastically dependent on the latent variables at the previous time point ($\boldsymbol{\theta}_{t-1}$). The latent variables ($\boldsymbol{\theta}_t$) reflect the mastery level of knowledge at each time point (Levy, 2019). Y. Han et al. (2022) combined dynamic Bayesian networks and IRT models to develop a sequential response model to estimate the continuous latent ability and the transitions of the observed problem states, namely the response sequences. More specifically, it described the probability of choosing the subsequent action given the individual's latent ability, the current action, and the tendency and correctness of the transition between successive actions. Similarly, Chen (2020) considered log events as a marked point process and proposed a probabilistic measurement model called the continuous-time dynamic choice model. The model described the probability of choosing the subsequent action given the entire event history and the respondent's ability. In addition, the model described the time associated with the next action given the event history and the respondent's speed. A further extension of the model is to incorporate an action-level easiness parameter into the action model (Xiao & Liu, 2023). As a special case of dynamic Bayesian networks, Hidden Markov Models (HMMs) have also been utilized to investigate the strategies employed by respondents (e.g. Arieli-Attali et al., 2019; Xiao et al., 2021). The authors considered an observed response to be determined by the current latent state of problem-solving and the previous action. Note that in HMMs, the latent state was assumed to have a discrete and finite state space.

In recent years, joint models of process data and performance data have also emerged. An important development in this area is the joint model of responses and response times. This model gained traction after van der Linden (2007) proposed a hierarchical framework. Since then, modeling responses and response times simultaneously has remained a vibrant research topic (Bolsinova & Tijmstra, 2018; Y. Liu et al., 2020; Zhan et al., 2023). Subsequent research has extended the concept to other facets of process data. For example, De Boeck and Scalise (2019) used task scores, time-on-task, and the number of actions as indicators for the latent variables performance factor, time factor, and action factor using confirmatory factor analysis (CFA; Jöreskog, 1969). Similarly, Lotz et al. (2022) extracted a process measure to determine the use of the VOTAT strategy for individual tasks. By aggregating the process indicators through factor analysis, they sought to examine whether the latent VOTAT variable moderated the relationships between intelligence and problem-solving competency.

## 2.3 Estimation of latent variable models

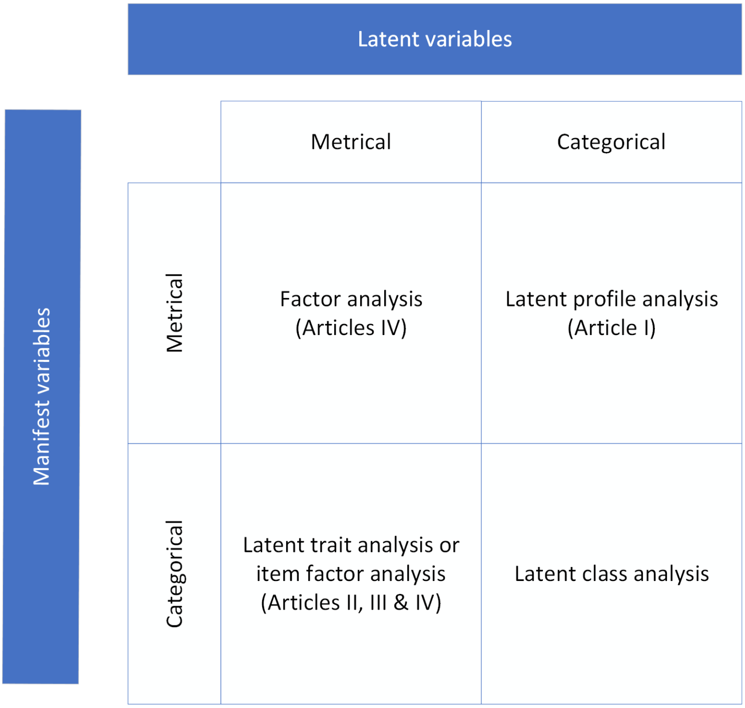
As introduced in the previous subsection, statistical modeling of process data can focus on observed indicators only or both observed indicators and latent constructs. Since many concepts in the project, such as problem-solving competency and cognitive processes, are not directly observable, we adopt the latter approach. Namely, in the dissertation, we model performance data and/or process data within the framework of latent variable modeling. However, the joint modeling of process data and performance data becomes complicated in terms of its estimation, and we aim to provide a computationally efficient estimation method in Articles III and IV. Accordingly, this section introduces latent variable models and focuses on the estimation of latent variable models for different types of observed indicators.

### 2.3.1 Introduction to latent variable models

*Latent variables* refer to hypothetical variables that cannot be directly observed. Factors and constructs are equivalent terms to latent variables in social sciences such as psychology and education. Latent variables, such as intelligence and personality, are primarily conceptual and difficult to measure directly. To infer the latent variable of interest, a common practice is to develop a battery of items grounded in a specific theoretical framework. It is assumed that the responses to these items can be primarily attributed to the latent variable. For example, the PIAAC 2012 PS-TRE domain was developed to measure problem-solving competency in technology-rich environments, and respondents' answers to the items were collected and regarded as *observed indicators*, also called *manifest variables*, for problem-solving competency.

In the analysis of the observed indicators, researchers extract what is common in the indicators. The latent variable that accounts for the common variability of the observed indicators is interpreted as problem-solving competency afterward. Similarly, other latent variables can be assumed to underlie the observed process indicators, such as speed underlying time-on-task (De Boeck & Scalise, 2019; van der Linden, 2007). In addition to the theoretical explanation of the response data, the latent variable approach is attractive because it reduces the dimensionality of multivariate data (Bartholomew et al., 2011). The purpose of dimension reduction is to use fewer variables to capture the same essential information embedded in the original variables. In this example, instead of focusing on the individual score of each item in the PS-TRE domain, we can use only the value of the latent variable to indicate the respondents' level of problem-solving competency.

After understanding the rationale for latent variables, the next step is to model the relationship between the latent variables and the observed indicators. Such models are called latent variable models. Based on the distributions of the observed and latent variables, latent variable models can be classified into different types. Let us classify random variables into two main categories: metrical variables whose realization falls in the set of real numbers and categorical variables whose value is one of a set of ordered or unordered categories. Based on this, latent variable models can be classified as factor analysis (metrical observed indicators and metrical latent variables), latent trait analysis/IRT (categorical observed indicators and metrical latent variables), latent profile analysis (metrical observed indicators and categorical latent variables), and latent class analysis (categorical observed indicators and categorical latent variables) (p.11; Bartholomew et al., 2011). An illustration is shown in Figure 2.5. We employed latent profile analysis in Article I to cluster respondents' solution patterns. In Articles II to IV, we assumed continuous latent variables with



| | | Latent variables | |
|---|---|---|---|
| | | Metrical | Categorical |
| Manifest variables | Metrical | Factor analysis<br>(Articles IV) | Latent profile analysis<br>(Article I) |
| | Categorical | Latent trait analysis or<br>item factor analysis<br>(Articles II, III & IV) | Latent class analysis |

Figure 2.5: Classification of latent variable models.

different types of observed indicators and applied CFA with different types of indicators. Below, I mainly introduce the basics and estimation of CFA for different data types.

### 2.3.2 Estimation of confirmatory factor analysis

In statistical modeling, we collect observed data from an unknown population that can be represented as a probability distribution, and our goal is then to identify the most likely population (i.e., the most likely probability distribution) that generated the observed data (Myung, 2003). A probability distribution contains model parameters, and therefore, we aim to find the values of the model parameters that best fit the observed data. This process is called parameter estimation.

Using a mathematical representation, let us denote $f(\boldsymbol{y}|\boldsymbol{\theta})$ as the probability or probability density function of observing the collected data $\boldsymbol{y}$, given a set of model parameters $\boldsymbol{\theta}$. Then, we aim to find the values of $\boldsymbol{\theta}$ that fit best with the observed data. Three general approaches are used to estimate the model parameters: least squares estimation (LSE), maximum likelihood estimation (MLE), and Bayesian methods. LSE aims to minimize the distance between the sample variance matrix and the model-implied covariance matrix (p.59; Bartholomew et al., 2011). Unlike LSE and MLE which regard the unknown parameters as fixed quantities, the Bayesian approach considers the parameters to be varied quantities that can be expressed with a probability distribution (p.324; Casella & Berger, 2021). Bayesian methods specify prior distributions of the parameters based on the belief of data analysts and update the prior distributions by including information from the collected sample. The updated priors are called the posterior distribution and Bayesian inference is based on the converged posterior distribution.

In this dissertation, we adopt maximum likelihood estimation and will only introduce MLE in the following sections. MLE has the advantage of sufficiency, consistency, efficiency, and parameterization invariance (Myung, 2003). In this approach, we need to write the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. It is a function of the model parameters and indicates the likelihood of the model parameters $\boldsymbol{\theta}$ conditional on the particular observed data $\boldsymbol{y}$. The exact formula needs to be updated according to a particular model specification. The next step is to maximize the (log) likelihood function such that the desired probability distribution makes the observed data most likely. To do so, we can make use of the partial differential function if the maximum likelihood estimates exist and are unique and the log-likelihood function is differentiable. Namely, we set $\frac{\partial logL(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\boldsymbol{\theta}} = \mathbf{0}$, denoted as the likelihood equation. To ensure that a maximum has been attained it is necessary to check that the shape of the log-likelihood function is convex. This can be done by verifying that the second-order derivatives are negative. However, it is often not possible to solve the equations analytically. In such cases, numerical optimization algorithms such as the Newton-Raphson or gradient descent algorithms are often used. These algorithms start at certain initial parameter values at random or by guessing and then iteratively update the estimates until fulfilling the stopping rules. Since $\boldsymbol{y}$ can have different variable types, the distribution function $f(\boldsymbol{y})$ and the likelihood function will need to adjust to the variable types.

Factor analysis can be categorized as exploratory and confirmatory. In this thesis, we focus exclusively on the confirmatory approach. The reason is that we have assumptions about the relationships between the observed and latent variables. CFA is particularly appropriate when researchers have hypotheses about the structure of the latent variables, namely which items are indicative of a particular factor. These hypotheses are often based on established theories or empirical studies. In the following subsections, I discuss the estimation of CFA with different types of indicators using a maximum likelihood estimator.

#### 2.3.2.1 CFA with continuous data

Here I introduce CFA with continuous indicators. Let $\boldsymbol{y}$ denote an $I$-dimensional vector of observed variables, $\boldsymbol{b}$ denote an $I$-dimensional vector of means or intercepts, $\boldsymbol{\Lambda}$ denote an $I \times P$ matrix of factor loadings, $\boldsymbol{z}$ denote a $P$-dimensional vector of latent variables, and $\boldsymbol{e}$ denote an $I$-dimensional vector of error terms.

Factor analysis can be expressed by the following formula,

$$\boldsymbol{y} = \boldsymbol{b} + \boldsymbol{\Lambda z} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$ and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. $\boldsymbol{\Psi}$ is a diagonal matrix, implying that the error items are independent of each other. This model suggests that the value of the observed indicator is determined by the average effect of a given item, the impact of the latent variables, and the uniqueness of the item. Note that the observed indicators $\boldsymbol{y}$ are assumed to be multivariate normally distributed and the dispersion matrix of $\boldsymbol{y}$ can then be expressed as:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}. \tag{2.2}$$

$\boldsymbol{\Sigma}$ denotes the theoretical population covariance matrix based on Equation 2.1. $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ are the model parameters that we aim to estimate. In a maximum likelihood approach, we write the log-likelihood function,

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \log \int \prod_{i=1}^{I} P_i(\boldsymbol{y}_i|\boldsymbol{z})\psi(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Phi})d\boldsymbol{z}, \tag{2.3}$$

where $\psi(\cdot)$ is the multivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Phi}$. It can be seen that Equation 2.3 integrates over the latent variables. This process is known as marginalization and thus referred to as marginal maximum likelihood (MML; Bock & Aitkin, 1981). After some operations, the log-likelihood function can be organized as (Jöreskog, 1969):

$$l = constant - \frac{N}{2}[log|\boldsymbol{\Sigma}| + trace(\boldsymbol{S\Sigma}^{-1})], \tag{2.4}$$

where $\boldsymbol{S} = \sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{b})(\boldsymbol{y}_i - \boldsymbol{b})'/N$, which directly summarizes the pairwise covariances between all pairs of observed data in the sample. To seek the model estimates that maximize the log-likelihood function, derivative functions are then needed. The details of the derivatives are omitted here due to the page limitations. Maximum likelihood estimations are widely implemented in software such as M*plus* (L. K. Muthén & Muthén, 2010) and the *lavaan* R package (Rosseel, 2012).

#### 2.3.2.2 CFA with ordinal data

When the responses are categorical data such as binary or ordinal data, Equation 2.1 cannot be applied directly. The multivariate normal distribution is severely violated if the observed data are discrete data with a few categories. As a result, Equation 2.2 does not hold in this situation. Applying MLE directly to factor analysis with ordinal data can result in biased factor loadings and standard errors and inflated chi-square statistics (C.-H. Li, 2016). To handle this situation, item factor analysis (IFA; Mislevy, 1986) or the latent trait model is proposed. IFA can fall into both the Underlying Response Variable framework

(URV; Jöreskog, 1994; B. Muthén, 1984) and the IRT framework (Samejima, 1969). Both approaches will be discussed here.

The URV approach assumes that there are continuous latent responses denoted as $\boldsymbol{y}^*$ underlying the observed categorical responses denoted as $\boldsymbol{y}$. Note that the latent responses $\boldsymbol{y}^*$ are different from the latent constructs $\boldsymbol{z}$. The latent constructs $\boldsymbol{z}$ dominate the latent responses $\boldsymbol{y}^*$. The latent responses are associated with items. For example, $y^*_{if}$ indicates the underlying responses of individual $f$ on item $i$. The latent responses connect to the observed data $y_{if}$ with $m_i$ categories through

$$y_{if} = c, \text{ if } \tau_{ic-1} < y^*_{if} < \tau_{ic}, \tag{2.5}$$

where $-\infty = \tau_{i0} < \tau_{i1} < \cdots < \tau_{im_i-1} < \tau_{im_i} = +\infty$ are the threshold parameters of item $i$. The latent responses $\boldsymbol{y}^*$ are often assumed to be standard normal distributed (Jöreskog, 1994). Under the URV approach, the probability of a response vector can be written as,

$$P(y_1 = c_1, y_2 = c_2, ..., y_I = c_I|\boldsymbol{\theta}) = \int_{\tau_{1c_1-1}}^{\tau_{1c_1}} \cdots \int_{\tau_{Ic_I-1}}^{\tau_{Ic_I}} \psi(\boldsymbol{y}^*; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{y}^*})d\boldsymbol{y}^*, \tag{2.6}$$

where $\psi(\cdot)$ is a $I$-dimensional normal density function with zero means and covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{y}^*} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{2.7}$$

Equation 2.6 requires computing $I$-dimensional integrals, which is difficult to obtain.

Under the URV framework, the parameters to be estimated are the thresholds, factor loadings, and the polychoric correlations - the correlations between two latent response variables for polytomous items. The estimation can be categorized into four main approaches: three-stage estimation methods (Jöreskog, 1994; B. Muthén, 1984), pairwise maximum likelihood (Katsikatsou et al., 2012), robust maximum likelihood (Yang-Wallentin et al., 2010), and full information maximum likelihood. In the three-stage estimation methods, thresholds are estimated at the first stage based on the univariate marginal likelihoods $l_{ic_i} = \int_{\tau_{ic_i-1}}^{\tau_{ic_i}} \psi(\boldsymbol{u})d\boldsymbol{u}$, where $\psi(\boldsymbol{u})$ denotes a standard normal distribution. Second, the polychoric correlations are estimated by maximizing the bivariate marginal likelihoods $l_{ic_i,jc_j}$ where $i > j$ given the estimated threshold parameters. Third, Equation 2.7 is fitted by imputing the results from the second stage to obtain the structural parameters like factor loadings using the generalized least square methods(B. Muthén, 1984) or the weighted least squares method (Jöreskog, 1994). The second approach, pairwise maximum likelihood, only considered each pair of items at one time and then sum over the pairwise log-likelihoods (Katsikatsou et al., 2012). The model estimates are obtained simultaneously when maximizing the summed pairwise log-likelihoods. This method is relatively fast because it only requires two-dimensional integrations. Third, robust maximum likelihood methods have been proposed to deal with data when the multivariate normal distribution assumption does not hold. Robust maximum likelihood methods correct the standard errors of the maximum likelihood estimator and the chi-square test statistics to reduce the impact of non-normality (C.-H. Li, 2016). The first three approaches only use part of the information from the observed data, such as the polychoric correlations, and thus falling into the category of limited-information estimation methods. In contrast, the fourth approach uses a full information maximum likelihood (FIML) estimator. This method seeks estimates that maximize Equation 2.7 by using the $I$-dimensional integral for all the possible response patterns. Because of the complexity of high-dimensional integrals, FIML is not feasible under the URV framework with more than five items (Jöreskog & Moustaki, 2001).

Another framework to model IFA is the IRT framework (Lord, 1980). IRT refers to a set of item response models that provide a probabilistic representation of the relationship between the observed categorical indicators and the latent trait through item characteristic curves describing the probability of correctly answering an item across the continuum of the latent trait, or item response category characteristic curves describing the probabilities for polytomous responses. Similar to the URV framework, IRT models assume the local independence of responses given the latent traits. It is worth noting that we can define the same model using the URV or IRT approach. This can be achieved by using a probit rather than a logit link function in IRT models. Both approaches can use MML estimation. An important difference between the approaches is that the estimation methods defined for the URV approach cannot be directly utilized to estimate the logit model.

A commonly used IRT model for binary data is a two-parameter logistic model:

$$P(y_i = 1|z) = \frac{1}{1 + \exp\left[-a_i(z - b_i)\right]}, \tag{2.8}$$

where $z$ is a latent trait being measured, $a_i$ and $b_i$ are the discrimination and difficulty parameters of item $i$. The discrimination parameter indicates the power of the item to distinguish lower and higher levels of respondents. The difficulty parameters describe the level of the latent trait that has 50% probability of endorsing item $i$. For polytomous data, the graded response model (Samejima, 1969) is often used:

$$P(y_i = c|z) = \frac{1}{1 + \exp\left[-a_i(z - b_{ic})\right]} - \frac{1}{1 + \exp\left[-a_i(z - b_{ic+1})\right]}, \tag{2.9}$$

where $b_{ic}$ indicates the boundaries between category $c$ and category $c + 1$. Note that the item parameter estimates of IFA under the URV framework and the IRT framework can be approximately transformed (Wirth & Edwards, 2007).

The estimation methods under the IRT framework generally rely on the raw data and make full use of the information in the data. Two main approaches are proposed for the estimation of IRT models. One is the joint maximum likelihood approach (JML; Wingersky, 1983). JML focuses on the log-likelihood function

$l(\boldsymbol{z}; \boldsymbol{a}, \boldsymbol{b})$ and seeks for the person and item parameters simultaneously that maximize the log-likelihood function. JML works best when the sample size is large (e.g., 1000 respondents) and the test is long (e.g., 60 items). If the data only correspond to a small number of test-takers and items, the model parameter estimates can be biased (Lord, 1986). JML has two main drawbacks. First, JML is computationally demanding because it estimates all the parameters at the same time. Second, JML has the issue of consistency, which means that, for a fixed number of items, increasing the sample size cannot guarantee better estimates. Instead of working on the joint likelihood function, MML integrates out the latent variables from the original log-likelihood function and then looks for item parameters that maximize the marginal likelihood function $l(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{z})$. The person parameter estimates are obtained subsequently given the item parameter estimates. MML reduces the computational burden compared to JML because the estimation of item parameters and person parameters are separated. However, when the latent variable vector has a high dimension, the integrals become too complicated and even infeasible. In practice, we do not compute the integrals directly; instead, we use numerical integration methods to approximate the high-dimensional integrals, such as quadrature-based integration, Bayesian methods, and Laplace approximations. The numerical integration methods will be introduced more in the Methods chapter.

#### 2.3.2.3 CFA with continuous and ordinal data

In psychology and educational measurement, it is also common to have a combination of continuous and discrete data. For example, background questionnaires often include questions about age, weight, and height (i.e., continuous variables) as well as four-point or seven-point Likert-type scales (i.e., ordinal data). Another example is that process data often consist of (ordinal) responses and (continuous) response times. To estimate CFA models with continuous and ordinal data, two frameworks can be applied: the URV (B. Muthén, 1984) framework and generalized linear latent variable models (GLLVMs; Bartholomew et al., 2011; Huber et al., 2004; Rabe-Hesketh et al., 2004).

The URV framework has been introduced in Section 2.3.2.2. The latent response variables $\boldsymbol{y}^*$ are assumed underlying the observed variables $\boldsymbol{y}$ (B. Muthén, 1984). For continuous data, the latent response variables can be directly observed, namely $\boldsymbol{y}^* = \boldsymbol{y}$. For ordinal observed variables, the latent responses are linked to the ordinal indicators via the thresholds (see Equation 2.5). The three-stage estimation method (Jöreskog, 1994; B. Muthén, 1984) can also be applied to this situation. Compared to analyzing ordinal data only, the three-stage estimation method involves not only polychoric correlations between ordinal data, but also polyserial correlations (Olsson et al., 1982) between ordinal data and continuous data and Pearson correlations between continuous data. Bayesian methods are also used to model a mixture of continuous and discrete data. For example, Bayesian methods have been used to jointly model responses and response times (e.g., Y. Liu et al., 2020).

Another framework is GLLVMs. As the name suggests, there are three keywords to understand GLLVMs. First, *generalized* means that the outcome variable can be generalized to bounded and/or non-continuous data, such as binary and count data. Second, *linear* refers to a linear combination of predictors. Third, *latent variable models* imply that latent variables are introduced to capture the common variance of a set of observed variables. Hence, the distributions of the observed variables are conditional on the latent variables. Instead of assuming a latent response variable underlying the observed variable, GLLVMs connect the linear combination including item parameters and the latent variables to the observed variable via a link function such as an identify, logit, or probit function. Since the latent variables are not observable, we need to integrate the latent variables out of the likelihood function. That is, the MML estimation approach will be used. As mentioned in the estimation within the IRT framework, the MML estimator requires complicated integrals and numerical approximation methods are required to approximate the MML. The details of the approximation methods will be introduced in Section 4.3.3.2.

Chapter 3

# Theoretical foundations for the empirical studies

This chapter discusses the theoretical foundations for the empirical studies presented in Articles I and II. There are two basic theoretical foundations: a) the problem-solving theory directly used in the dissertation, and b) the rationale behind process data analysis.

## 3.1 Mayer and Wittrock's problem-solving theory

Although there have emerged many problem-solving theories, the theory proposed by Mayer and Wittrock (1996, 2006) is one of the most widely used theories in problem-solving research so far. Their problem-solving theory also lays the theoretical foundation for this dissertation. Their theory had a domain-specific focus, but it has been extended to domain-general problem-solving as well in the PISA and PIAAC assessments. The theory of Mayer and Wittrock (2006) consists of four cognitive processes, including *representing*, *planning/monitoring*, *executing*, and *self-regulating*.

*Representing* refers to the process of converting a problem into a mental representation. This involves using, for example, symbols, equations, and graphs to formulate the problem situation (OECD, 2014b). Relevant to information-processing theory, the representing process involves the construction of a problem space (Simon & Newell, 1971). To get a better understanding of the problem, problem-solvers might need to interact with the problem environment by conducting activity-based manipulations (Jonassen, 2000). Compared with novices, experts perform better in recognizing the problem space (Jonassen, 2010).

*Planning* refers to devising a method to solve a problem such as decomposing an overarching goal into several subgoals, and *monitoring* describes the process of evaluating whether the plan is appropriate and effective. The key to planning is to construct a future action sequence and organize activities to fulfill certain outcomes through deliberate and thoughtful considerations (Mumford et al., 2001). Monitoring involves reflections on the conceived plans and relates to the skills of foreseeing potential issues of the plans.

*Executing* refers to the process of carrying out the plan, which can be observed directly. In this process, problem-solvers engage in a series of actions or operations, such as writing down a formula to solve a math problem and clicking on the buttons on a screen to solve problems presented by computers. In fact, log files recorded by computers directly reflect the executing process.

*Self-regulating* involves the modification or maintenance of cognitive activities moving toward the goal. For example, if a problem-solver applied a plan but failed to solve the problem, then this individual may restart the task to adopt a new solution or revise certain operations that seemed erroneous.

The framework proposed by Mayer and Wittrock (2006) is relatively general and applicable to many types of problems. The broad acceptance of this framework implies its significance in problem-solving research. This framework is related to concepts such as the problem space from the information-processing approach and shares great similarities with the theory of Polya (2004) from a mathematical problem-solving perspective. Because various theories use different terms to describe their problem-solving models, the PISA and PIAAC expert groups summarized and organized the cognitive dimensions of problem-solving. They developed their frameworks for the problem-solving domain by synthesizing the existing theories, particularly those by Mayer and Wittrock (2006) and Simon and Newell (1971). The details of the framework in PISA and PIAAC are introduced in Section 2.1.4.5. On closer inspection, PISA 2012 (OECD, 2014b) stresses the representing process, specifically highlighting the process of exploring and understanding the problem. In comparison, PIAAC 2012 (OECD, 2012) places a significant emphasis on "using information". This stems from its use of information items, which rely heavily on the application of information for problem-solving. Additionally, it is evident that the frameworks are developed and modified according to their focus in each cycle. For example, the collaborative problem-solving dimensions are added into the framework in PISA 2015 compared to PISA 2012 (OECD, 2017). PIAAC also shifts from PS-TRE in the first cycle to adaptive problem-solving in the second cycle (Greiff et al., 2017). This dissertation intends to use a fundamental and comprehensive framework of cognitive processes in problem-solving that shares the essential components of cognitive processes with the theoretical frameworks of both PISA and PIAAC.

## 3.2 Theoretical foundations for process data analysis

The rationale behind the use of process data constitutes another central theoretical foundation for this dissertation. We aim to gain a deeper understanding of problem-solving utilizing insights from process data. This section presents why process data can shed light on problem-solving processes and provide evidence of validity.

As defined by Mayer and Wittrock (2006), problem-solving refers to cognitive processes when the solution is not obvious, which means that problem-solving involves sequential steps (Zoanetti & Griffin, 2017). For example, an early stage of problem-solving often involves understanding the problem and representing the problem in mind (Mayer & Wittrock, 1996; OECD, 2012, 2014b; Polya, 2004). This means that respondents may need to spend more time reading the instructions or navigating through different interfaces of the task. By reviewing individual process data, it is possible to infer if a respondent rushed to interact with the problem environment (e.g., the time to the first action was too short to read through the instructions) or a respondent spent an extended time reading the instruction or making a plan before taking actions (e.g., the time to the first action was quite long compared to the subsequent executing process). Namely, process data can potentially serve as a window into the black box of the mental processes of problem-solvers (Bunderson et al., 1989; Greiff et al., 2015). If we can identify the strengths and weaknesses of test-takers in terms of different stages of problem-solving, it is possible to provide a more fine-grained diagnosis of problem-solving competency (Zoanetti & Griffin, 2017) or more personalized and real-time feedback for different test-takers during the assessment to aid problem-based learning.

In addition, process data describe the response process that provides evidence of validity - the extent to which the interpretation of test scores can be supported by evidence or theory (AERA, 2014). To be specific, response processes are a source of evidence of validity in terms of the fit between the conceptual construct (e.g., problem-solving competency) and the actual and detailed nature of the performance (AERA, 2014). For example, the PISA 2012 *Climate Control* task has a pre-defined optimal exploration strategy - VOTAT, and the task's validity can be assessed by examining if the use of the VOTAT strategy benefits task performance. In line with this thinking, Greiff et al. (2015) analyzed process data to evaluate the degree to which students actually applied VOTAT and its relationships with task performance. That is, process data can help interpret test scores and enrich the definition of the construct (AERA, 2014). Besides providing construct-relevant interpretations, process data can also provide evidence that test scores reflect something construct-irrelevant. For example, if a respondent did not interact with the problem environment but skipped the task (indicated by very few interactions with the computer) or gave a random answer quickly (indicated by a short response time and lack of necessary actions), it seems more suitable to interpret such behavior as disengagement rather than low problem-solving competency. For test developers, it is particularly crucial to have a clear definition of the construct being measured (e.g., problem-solving competency) and expectations of respondents' behavior. Comparing the actual behavior of test-takers and the expected behavior can help test developers validate the task design (Zoanetti & Griffin, 2017). In summary, understanding and making use of process data can contribute to the validation of assessments, inference of unobserved mental processes, and the design of personalized tools to improve problem-solving competencies.

## 3.3 Relationships between theoretical foundations and articles

Figure 3.1 displays the relationships between the theoretical foundations for the substantive perspective of the dissertation and the first two articles. Both articles are rooted in the problem-solving theory of Mayer and Wittrock (2006) and the rationale for using process data.

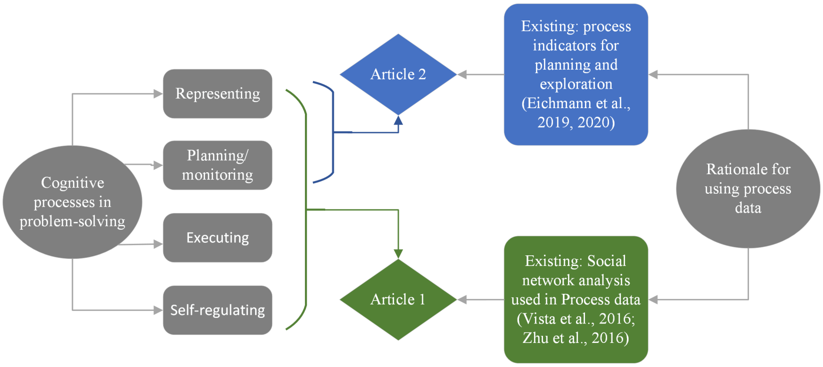

Cognitive processes in problem-solving
Representing
Planning/ monitoring
Executing
Self-regulating
Article 2
Article 1
Existing: process indicators for planning and exploration (Eichmann et al., 2019, 2020)
Existing: Social network analysis used in Process data (Vista et al., 2016; Zhu et al., 2016)
Rationale for using process data

Figure 3.1: The theoretical framework of Articles I and II.

### 3.3.1 From theoretical foundations to Article I

In Article I, we aim to visualize and understand the solution patterns of the respondents. To achieve this goal, we employ social network analysis (Newman, 2018) because it can reflect the actual activities performed by the problem-solvers and be combined with the problem-solving theory to determine the network features to extract.

There are primarily three existing articles that provide us with a more relevant basis for Article I. Vista et al. (2016), Vista et al. (2017) and Zhu et al. (2016) utilized directed networks to represent action sequences with vertices representing single actions and directed edges representing the transitions from one action to another action. Vista et al. (2016), Vista et al. (2017) aimed at finding the prominent interaction sequence based on the importance of the vertices and edges. Specifically, the authors computed the *eigenvector centrality* metric (Borgatti, 2005) for each vertex (i.e., action in this example). For the edges, the importance was weighted by their frequencies. Namely, the importance of edges was defined by the frequency of transitions between the two actions. After filtering out less important vertices and edges, the authors then further explored the features of the remaining paths. Similarly, Zhu et al. (2016) also created a weighted directed network of action sequences using the frequency of transitions as the weight of edges. Different from Vista and coauthors, Zhu et al. (2016) extracted other features from the networks: *weighted density* to capture the average frequency of action transitions, *degree centrality* to indicate the frequency of each action performed, *reciprocity* to measure the tendency to immediately revisit previous actions, 16 *triadic patterns* to exhaustively describe the potential relationships between any arbitrary sets of three actions. After extracting these network features, the authors then used discrimination analysis to find the features that can significantly predict task performance scores based on certain rules.

These studies lay the foundation for Article I of this dissertation by providing the rationale and application of utilizing social network analysis to analyze process data. However, there are two main limitations: a) their studies used only action sequences but ignored the timestamps, and b) the extraction of network features was not based on a substantive framework. To overcome the limitations, Article I uses information from both action sequences and response times and extracts network features guided by the problem-solving theory (Mayer & Wittrock, 2006). Rather than identifying the prominent action sequences or examining the predictive power of network features, we identify typical solution patterns with seven network features that we extract from action sequences and response times. The number of solution patterns is determined by an exhaustive comparison of models with one to nine clusters. That is, Article I combines both theory-driven and data-driven approaches. A task from the PISA 2012 problem-solving domain is used to demonstrate the proposed method.

### 3.3.2 From theoretical foundations to Article II

In Article II, we focus on the cognitive processes of representing and planning. These unobserved cognitive processes are inferred by process indicators extracted from process data using a theory-driven approach. Existing studies have defined process indicators to reflect the planning and representing processes, which are briefly summarized below.

In the literature, both qualitative and quantitative methods have been applied to infer planning - mental simulations of future operations and associated outcomes (Mumford et al., 2001). Planning is resource-intensive and requires a relatively long time compared to executing the plan. Most empirical studies that relate to planning have focused on static problems. A commonly used measure of planning in static problems like the Tower of London is the first-move latency (e.g., Albert & Steinberg, 2011; Unterrainer et al., 2003). It is defined as the time interval between the start of a problem and the first operation performed by a respondent. This measure has been found to be positively related to task performance (Albert & Steinberg, 2011).

However, in dynamic problems, the first-move latency may not be so appropriate because respondents may need to interact with the computer to perform an environment analysis to gather enough information before making a plan. Therefore, Eichmann et al. (2019) proposed to use the longest duration indicator (defined as the longest time interval between two successive actions) to represent the planning process. Note that the longest duration can occur at any time when solving the problem. In addition, Eichmann et al. (2019) also proposed two other process indicators: the delay indicator (the time taken before the longest duration appears) and the variance indicator (the variation of all durations). The authors then used the three indicators to predict task success and found a non-significant effect for the longest duration indicator and the variance indicator but a negative effect for the delay indicator and some interaction effects of the delay indicator with other indicators. This suggests that early planning benefits task performance, but continuous planning or extended longest duration can compensate for the lack of early planning (Eichmann et al., 2019). However, two critical questions arise: a) Do the indicators from different tasks imply the same cognitive process of planning? and b) would the relationships between the planning indicators and task performance vary across tasks due to different task characteristics such as interfaces and complexities?

Another cognitive process that we focus on is representing. In order to better represent dynamic problems, respondents need to interact with the computer because not all the necessary information is presented at the outset. Interactions can be categorized into goal-directed behavior or non-targeted exploration (Eichmann, Goldhammer, Greiff, et al., 2020; Eichmann, Greiff, et al., 2020). Goal-directed behavior refers to the operations that are required to solve the problem, whereas non-targeted exploration refers to the operations that are not necessary to solve the problem (i.e., not included in optimal solutions), such as checking the help menu or resetting the task. It can be seen that goal-directed behavior conveys similar information as task success, whereas non-targeted exploration appears erroneous. However, it has been documented that non-targeted exploration benefits task performance (Dormann & Frese, 1994) and meta-cognition (Bell & Kozlowski, 2008), including planning and self-monitoring. An explanation is that non-targeted exploration assists in representing the problem and getting a better understanding of the available features of the problem environment (Eichmann, Greiff, et al., 2020).

Another aspect of an action is if it is conducted for the first time. If an individual performs a non-targeted operation for the first time, the relevant information or functions of the operation can be incorporated into the problem space; namely, it relates to information generation (Wüstenberg et al., 2012). If the same action is performed by the individual again, it could not bring new information and indicates "an overestimation of the relevance of the inspected information" (Eichmann, Greiff, et al., 2020). Eichmann, Greiff, et al. (2020) labeled the operations with initial/repeated non-targeted exploration or initial/repeated goal-directed behavior and applied a full-path sequence analysis based on these categories of actions to profile students. In the study of Eichmann, Goldhammer, Greiff, et al. (2020), initial and repeated non-targeted exploration were not distinguished and the number of non-targeted exploration was computed by the Levenshtein distance between the shortest correct solution path (i.e., optimal solutions) and the observed action sequence. Although non-targeted exploration was viewed as a latent variable and functioned as a moderator variable between gender/immigrant background and problem-solving competency, the model fit of the non-targeted exploration model was not reported in their papers.

It can be seen that the above-mentioned empirical studies have extracted process indicators from the response processes to reflect the unobserved cognitive processes. However, the construct validity of the process indicators defined in these studies has not been explicitly examined. To be specific, to what extent can we interpret the longest duration as a measure of planning (Eichmann et al., 2019), and to what extent can we interpret the number of non-targeted behaviors as an indicator of non-targeted exploration (Eichmann, Goldhammer, Greiff, et al., 2020; Eichmann, Greiff, et al., 2020)? Article II aims to test the construct validity of the indicators for planning and non-targeted exploration based on the work of Eichmann. It additionally investigates the relationships between problem-solving competency, planning, and non-targeted exploration at both an overall and task-specific level.

Chapter 4

# Methods and methodological reflections

The dissertation consists of four articles, and all of them utilize quantitative research methods. Quantitative research methods involve quantifying numerical data and analyzing the collected data using statistical methods for testing hypotheses and examining relationships (Apuke, 2017). This chapter describes the elements of the quantitative research methods used in the dissertation. These include where we get the data (Section 4.1), how we define the numerical measures (Section 4.2), and what methods are employed to answer our research questions (Section 4.3). Table 4.1 outlines the methodological elements for Articles I to IV, including the perspectives (substantive or methodological), the type of studies (empirical or simulation), the source of data, data preprocessing including data cleaning and recoding, the measures defined from the process data, the statistical models, and the software used to implement the models. In addition, a short discussion of some ethical considerations is provided in Section 4.4.

## 4.1 Data

This dissertation uses data from international large-scale assessments; namely, we are conducting secondary data analysis (Johnston, 2014). International large-scale assessments are conducted by intergovernmental organizations (e.g., OECD) across countries/economies to monitor the trends in cognitive skills of interest and provide comparative results across participating countries/economies, having important implications for policymakers. Moreover, these assessments typically collect demographic information about the participants, which can be used to explain individual differences in cognitive skills, for example. These assessments provide a rich source of data for researchers, and the results provide valuable insights for educators and policymakers. This dissertation uses data from PISA and PIAAC. Both PISA and PIAAC have taken place in many countries/economies, providing valuable insights into country-level performances and comparisons among countries. Moreover, they are both repeated cross-sectional studies that provide the trend of the skills over time by linking the scales from different rounds using standard common item equating methods (OECD, 2014b). Although differing in target populations and specific assessment designs, PISA and PIAAC share similar conceptual frameworks of constructs and methods of assessment (OECD, 2021). Below is a brief description of PISA and PIAAC.

The overarching goal of PISA is "to measure how well 15-year-old students approaching the end of compulsory schooling are prepared to meet the challenges

Table 4.1: A summary of methodological elements in the dissertation.

| Element | Article I | Article II | Article III | Article IV |
|---|---|---|---|---|
| Perspective | • Substantive | • Substantive | • Methodological | • Methodological |
| Aims | • Define networks from action sequences and response times<br>• Extract network features based on problem-solving theory<br>• Identify solution patterns based on network features | • Examine the construct validity of planning and non-targeted exploration indicators<br>• Test the relationships between planning, non-targeted exploration, and problem-solving competency | • Develop a fast estimation method based on a second-order Laplace approximation to estimate GLLVMs with categorical data | • Apply a fast estimation method based on Laplace approximations to estimate GLLVMs for ordinal, continuous and count data, which can be applied to joint models of performance data and process indicators |
| Type | • Empirical | • Empirical | • Simulation & Empirical | • Simulation & Empirical |
| Data | • *Traffic* task from the PISA 2012 creative problem solving domain<br>• Log files of 406 students from the USA | • Seven tasks from the PIAAC 2012 PS-TRE domain<br>• Log files of 1325 participants from the USA | • Four simulation conditions with 1000 replications for each condition<br>• Cognitive responses of 21672 students in PISA 2009 | • Two simulations with 8/16 conditions; 1000 replications for each condition<br>• Log files of 1029 students from Australia in the PISA 2012 mathematics domain |
| Data preprocessing | • Recode "hit_path" as "select_path"/"cancel_path"<br>• Recode actions beyond task instructions as an "insignificant" action | • Recode process indicators as categorical variables using quartiles as cutoff values | • Remove respondents with excessive missing values | • Log-transform and center response times<br>• Exclude outliers for process indicators |
| Measures | • Seven network features | • Response scores, planning and non-targeted exploration indicators | • Response scores | • Response scores, response times, and the number of actions |
| Methods | • Social network analysis for visualization and feature extraction<br>• GMMs for clustering students and identifying solution patterns | • Unidimensional CFA with ordinal data<br>• Multidimensional CFA with residual correlations considered | • Implement Lap2 efficiently<br>• Compare Lap1, Lap2, and AGHQ methods using simulations | • Implement Lap1 and Lap2 based on data types<br>• Compare Lap1 and Lap2 in estimating GLLVMs |
| Packages | • *igraph* and *mclust* | • *lavaan* | • *lamle* | • *lamle* and C++ |

of today's knowledge societies" (OECD, 2014c). PISA surveys have taken place every three years since 2000 with participation from over 40 countries/economies. The surveys consist of two parts: a) background questionnaires, including Student Questionnaires, School Questionnaires, and Parent Questionnaires, and b) cognitive items. Cognitive items cover the domains of reading, mathematics, and science in each round. Some other domains such as problem-solving and financial literacy have been considered in certain rounds. We use PISA data in Articles I, III, and IV as Table 4.1 indicated. As introduced in Section 3.1, the theoretical framework of the PISA 2012 creative problem-solving domain shares great similarities with the problem-solving theory by Mayer and Wittrock (2006) that we use in this dissertation. Hence, it is reasonable to choose the PISA data for our empirical studies of investigating problem-solving behavioral patterns and cognitive processes. Although there are many items and a great number of samples, PISA has only released a limited number of log files. We select only a few tasks and samples in the articles to illustrate our approaches. This is mainly due to the availability of data and datasets with a large sample size.

Unlike PISA's target population of 15-year-old students, PIAAC is a global assessment of adult skills, including literacy, numeracy, and problem-solving in technology-rich environments, which are considered as a basis for "effective and successful participation in the social and economic life of advanced economies" (OECD, 2012). As rapid technological changes continue, it is increasingly critical to improve the skills relevant to understanding, interpreting, analyzing, and communicating complex information. Therefore, PIAAC 2012 particularly emphasizes Information and Communication Technology (ICT) skills in information items. The target population of PIAAC is "non-institutionalised adults aged 16-65 years normally resident in the national territory of the participating country" (OECD, 2021). The institution here means, for example, nursing homes, jails, and mental hospitals. PIAAC surveys consist of background questionnaires, reports on skills used in the workplace, and assessments of literacy, numeracy, and problem-solving skills (OECD, 2019). Their theoretical framework of PS-TRE is also similar to the theory of Mayer and Wittrock (2006), providing the rationale for connecting the PIAAC data to the fundamental problem-solving theory used in the dissertation. The first cycle of PIAAC took place in three rounds: 2011-2012 (24 countries), 2014-2015 (9 countries), and 2017-2018 (6 countries). The second cycle of PIAAC redeveloped the theoretical framework of assessments according to the contemporary understanding of the skills and the results from the first cycle (OECD, 2021). The data collection of the second cycle of PIAAC is still undergoing (2022-2023) and the results are not available so far (date until June 2023).

## 4.2 Measures

This section summarizes how the measures utilized in the dissertation are defined and the context in which they are obtained. In general, the measures used in the dissertation are defined based on process data, including action sequences and response times, to reflect valuable information about human-computer interactions. There have been many approaches for defining such measures (see Section 2.2.4.1). In this dissertation, we apply a network analysis approach to obtain network features as process measures/indicators in Article I and a theory-based approach to obtain measures in Articles II to IV. The measures defined in the articles are introduced in the following subsections.

### 4.2.1 Article I: Network measures

In Article I, we apply a social network approach to visualize process data and extract measures that reflect essential characteristics of the response process. The rationale for employing this approach is that the operations are not isolated but directly connected to the previous and next operations, and networks seem to be a feasible tool to present the transitive relationships (Vista et al., 2016; Vista et al., 2017; Zhu et al., 2016). Specifically, Article I defines process data (i.e., action sequences and response times) as weighted directed networks with vertices representing individual actions, directed edges representing transitions between two actions, and the weight of the edges reflecting the response time on the associated transitions. If the transitions are performed more than once, the average time is utilized as the weight. The time-weighted method for edges distinguishes our study from previous papers that use the frequency of transitions as the weight of edges (Vista et al., 2016; Vista et al., 2017; Zhu et al., 2016). In network graphs, the weight is reflected by the thickness of the edges. An example network graph of process data is given in Figure 3 of Article I.

Next, we set out to define network features reflecting the cognitive problem-solving processes (Mayer & Wittrock, 2006). First, exploration plays an important role in the *representing* process (OECD, 2014b), and the exploratory behaviors can be reflected by the vertices. Therefore, we define the vertex feature *operation diversity* to indicate the extent to which a respondent performs diverse actions to explore the problem environment. Second, we define two time-related measures to reflect the *planning/monitoring* process (Eichmann et al., 2019): *average time* and *standard deviation of time*. Subsequently, the key to the *executing* process is the order/transitions of actions (Mayer & Wittrock, 2006). Since the transitions of actions are represented by the edges in networks, we define edge features to reflect the sequential actions at an overall level (*edge density*), between two actions (*reciprocity*), among three actions (*transitivity*), and between the correct transitions and incorrect transitions (the *External-Internal* index). The *reciprocity* and *transitivity* features indicate if respondents tend to revisit previous operations and reflect the *self-regulating* process. The *External-Internal* index implies to what extent respondents constantly conduct correct operations. The detailed definition and calculation of the seven network features can be found in Table 2 of Article I.

### 4.2.2 Article II: Planning and non-targeted exploration indicators

In Article II, we focus on the cognitive process of *representing* and *planning* (Mayer & Wittrock, 2006). Process indicators for planning and non-targeted exploration have been defined in the previous studies (Eichmann, Goldhammer, Greiff, et al., 2020; Eichmann et al., 2019; Eichmann, Greiff, et al., 2020). Based on these studies, we modify the indicators for planning and non-targeted exploration, examine the internal construct validity of the indicators, and model the overall and task-specific relationships between problem-solving competency, planning, and non-targeted exploration.

As introduced in Section 3.3.2, Eichmann et al. (2019) proposed to use the longest duration indicator, the delay indicator, and the variance indicator to reflect planning. In Article II, we use only the longest duration indicator because the three indicators are highly correlated in our data, meaning that they convey very similar information regarding planning. The longest duration indicator is defined as follows in our study: the time intervals between consecutive events, excluding the time intervals for the last two events (i.e., "NEXT_INQUIRY" and "END"). This is because the last two actions indicate exiting from the task, and the time on these action transitions is more relevant to the reflection on the task, rather than making a plan before executing. Note that in the study of Eichmann et al. (2019), any time interval can be considered a planning indicator as long as it is the longest, no matter where it occurs. To identify the longest duration, we compute all the time intervals between two successive actions and find the longest one among these durations except for the last two durations.

To reflect the *representing* process, we follow the definition by Eichmann, Greiff, et al. (2020) and use only the number of initial non-targeted operations as the indicator for non-targeted exploration. This is because the initial operations are related to information generation, whereas repeated operations are related to information integration, which would not contribute to expanding the problem space (Wüstenberg et al., 2012). To compute the number of initial non-targeted explorations, we first need to define goal-directed behaviors and non-targeted operations by comparing each operation to the optimal solutions. Operations that occur in any of the optimal solutions are categorized as goal-directed, while the rest are non-targeted. Then we count the number of first-occurred, non-targeted operations as the indicator for non-targeted exploration.

It is worth noting that the planning indicators are continuous variables and highly positively skewed, and the non-targeted exploration indicators are count variables ranging from zero to hundreds. Namely, the process indicators deviate from the normal distribution, making it inappropriate to apply the conventional CFA (Jöreskog, 1969). We then convert the process indicators into ordinal data and apply item factor analysis. Specifically, we recode the two process indicators into equal-sized categories and treat them as ordinal data. This recoding method ensures that each category has a similar number of observations and avoids highly unbalanced frequencies of the categories.

### 4.2.3 Article III: Response scores

Article III aims to overcome the computational challenge posed by the high dimensionality of latent variables. Specifically, we propose a fast estimation method of multidimensional GLLVMs with multiple groups for categorical observed data using second-order Laplace approximations of the integrals in the marginal log-likelihood function. In particular, ordered categorical observed variables are of interest in this article.

To test the performance of the proposed method, we conduct a simulation study by generating samples and true model parameters ourselves under the framework of GLLVMs. The measures that we generate are binary responses of three or four items per dimension because previous studies have shown that Laplace approximations perform the worst in this situation. In addition, we employ a real dataset from PISA 2009 to illustrate our approach. We use 188 items from the domains of mathematics, reading, and science and 21672 respondents from Hong Kong, Macao, Shanghai, and Chinese Taipei. Among the 188 items, 12 items are scored in three categories and the remaining items are scored in two categories. More detailed information can be found in Section 4 of Article III.

### 4.2.4 Article IV: Response scores, response times, and the number of actions

We extend the estimation approach of Article III to Article IV by incorporating a collection of ordinal, continuous, and count observed variables into GLLVMs. This extension can be applied to the joint analysis of performance data and process data where different types of variables occur in a single model. To efficiently estimate such models, we apply the first- and second-order Laplace approximations to the integrals of the marginal log-likelihood function. We conduct both an empirical study and simulations to examine the performance of the proposed method.

The empirical study uses a task from the computer-based assessment of mathematics (CBAM) in PISA 2012. The task consists of three subtasks, and three observed variables are derived from each subtask: task scores (two dichotomous scores and a three-categorical score), response times (i.e., the time interval between entering and exiting the task), and the number of actions (i.e., counting how many interactions a respondent conducts in the computer environment). Response times are log-transformed to deal with the highly-skewed issue, and outliers of the process indicators are excluded. Note that instead of recoding the process indicators into categorical data in Article II, we treat the observed indicators as they are in Article IV. The results of the model parameters and the PISA item pools are then used to help us determine the range of the true parameter values in the simulation studies. The simulations generate three correlated latent variables, with ordinal variables (three categories), continuous variables (normal distributions), and count variables (Poisson or negative-binomial distributions) as indicators, respectively. We consider three or six items per dimension. The details of the simulation design can be found in the Section "Simulation study" in Article IV.

## 4.3 Latent variable models

Latent variables are assumed to underlie the observed indicators and can explain the common variance among the indicators that measure the same concept. Latent variable models analyze the relationships between the observed indicators and the latent variables and reduce the dimension of multivariate analysis. In this dissertation, latent variable models are involved in all the studies. I introduce the specific latent variable models that we use in each article in the following subsections.

### 4.3.1 Article I: Gaussian Mixture Models

In Article I, we aim to identify the solution patterns based on the network features extracted. We assume that there are latent subgroups of the respondents. The respondents from the same subgroup share a common solution pattern but show a distinct pattern compared to other subgroups (James et al., 2013). To find the latent groups, we apply Gaussian Mixture Models (GMMs; Fraley & Raftery, 2002). In GMMs, the latent variable is discrete rather than continuous. A multivariate Gaussian distribution is assumed for the indicators in each latent subgroup. Because we do not have sufficient prior knowledge about the number of subgroups, we run a number of possible GMMs with different numbers of subgroups and the features of the covariance matrix of the indicators and decide on the final model according to information criteria and the bootstrap likelihood ratio test. After determining the subgroups of respondents, we then aggregate the process data of respondents from each subgroup, construct subgroup-level network graphs, and interpret the results from the subgroups as typical solution patterns. In the article, we apply the proposed method to a creative problem-solving task from PISA 2012.

### 4.3.2 Article II: CFA with ordinal data

In Article II, we apply CFA (Jöreskog, 1969) to examine the construct validity of the indicators for planning, non-targeted exploration, and problem-solving competency. Three unidimensional CFAs are conducted to test three hypotheses: a) the latent variable planning underlies the longest duration indicator, b) the latent variable non-targeted exploration underlies the number of initial non-targeted operations, and c) the latent variable problem-solving competency underlies the response scores. Since all the indicators are ordinal data and the latent variables are assumed to be continuous, the analysis falls into latent trait analysis or item factor analysis. As discussed in Section 2.3.2.2, there are two main approaches for parameter estimation for item factor analysis: The URV approach and the IRT approach. We use *lavaan* R package (Rosseel, 2012), which employs the URV approach with the diagonally weighted least squares

(DWLS) estimator. We report the model fit indices to evaluate if the models fit the data, as well as the factor loadings to show the influences of the latent variables on the indicators.

In addition, we investigate the relationships between planning, non-targeted exploration, and problem-solving competency by placing the three aspects into a single model. The overall relationships among them are reflected by the estimated covariance of the latent variables. After considering the overall relationships, we also add residual correlations between indicators from the same item to reflect task-specific relationships. These task-specific relationships capture the unique relationship between the indicators from a given item.

### 4.3.3 Article III & IV: GLLVMs

In Articles III and IV, we focus on the estimation of GLLVMs. GLLVMs can be viewed as a combination of generalized linear models (GLMs; Nelder & Wedderburn, 1972) and confirmatory factor analysis (Jöreskog, 1969). GLMs extend ordinary linear regression models by allowing response variables to be from an exponential dispersion family, such as Gaussian and Poisson distributions, and by connecting linear predictors to the observed responses with a link function. In GLMs, there is only one response variable (outcome variable) and all the variables are observable. However, the questionnaires and scales in psychology and education usually consist of multiple related items measuring the same construct. Therefore, the dependencies among items should be considered, and latent variables can be introduced to account for item dependencies, as CFA does. By combining GLMs and CFA, GLLVMs can a) deal with various types of response variables including continuous and discrete data, b) include a combination of linear predictors, and c) use latent variables to explain the common variance of the response variables. GLLVMs are flexible, but the estimation for high-dimensional GLLVMs is computationally demanding, which limits the application of GLLVMs. In Articles III and IV, we apply a fast estimation method to deal with this problem. In the following, I introduce the mathematical representations of GLLVMs and the estimation methods.

#### 4.3.3.1 Specification of GLLVMs

Let $y_{if}$ denotes the observed response of individual $f \in \{1, \cdots, F\}$ on item $i \in \{1, \cdots, I\}$, $\boldsymbol{w}$ denote $D$-dimensional predictors, and $\boldsymbol{z}$ denote a $P$-dimensional latent variable vector that is assumed to follow a multivariate normal distribution. Following Rabe-Hesketh et al. (2004), a general formula for GLLVMs can be written as

$$g_i(E[y_{if}|\boldsymbol{w}, \boldsymbol{z}]) = b_i + \boldsymbol{\beta}_i^{'}\boldsymbol{w} + \boldsymbol{a}_i^{'}\boldsymbol{z}, \tag{4.1}$$

where $b_i$ is the intercept parameter of item $i$, $\boldsymbol{a}_i$ is a slope parameter or factor loading vector of item $i$, and $\boldsymbol{\beta}_i$ is a $D$-dimensional vector of regression coefficients. Given $\boldsymbol{w}$ and $\boldsymbol{z}$, the observed indicator $y_{if}$ is linked to the right-hand of Equation 4.1 via a link function $g_i$. The choice of a link function depends on the distribution of $y_{if}$. The measurement models for ordinal data (graded response model), continuous data (normal distributions), and count data (Poisson distributions or negative-binomial distributions) and the associated link functions are given in the method section of Article IV.

Let $\boldsymbol{y}$ be an $I$-dimensional response vector. The marginal log-likelihood can be expressed as:

$$l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w}) = \log \int \prod_{i=1}^{I} P_i(\boldsymbol{y}_i|\boldsymbol{w}, \boldsymbol{z})\psi(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{z}, \tag{4.2}$$

where $\boldsymbol{\theta}$ represent the unknown parameters, $P_i$ defines the measurement model for the response of item $i$, and $\psi(\cdot)$ denotes the multivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Equation 4.2 is a marginal log-likelihood function that includes a marginalization process, namely an integration over the latent variable vector $\boldsymbol{z}$. That is, the person parameters are integrated out of the function, and we then seek item parameters maximizing the marginal log-likelihood function. Directly calculating the likelihood in Equation 4.2 is not possible because the integrals do not have closed-form solutions for GLLVMs. In practice, numerical integration methods are commonly applied to approximate the integrals. The following subsection summarizes some approximation approaches.

#### 4.3.3.2 Approximation approaches

For the numerical integration methods, I introduce three approaches: a quadrature-based approach, a simulation-based approach, and Laplace approximations.

*1. Quadrature-based approach.* An example likelihood plot for a one-dimensional parameter is presented in Figure 4.1. The area under the likelihood curve cannot be calculated directly from a known formula, so it needs to be approximated. A standard method is to use quadrature-based integration. This approach divides the irregular area into several rectangles and sums the areas of the rectangles with a quadrature weight function to approximate the original area. More rectangles provide a more fine-grained approximation but with an increased computational burden. This gives a simple concept of quadrature-based methods, but the actual implementation is more complicated. Gauss-Hermite quadrature (GHQ) has been applied to estimate generalized latent trait models (Moustaki & Knott, 2000). GHQ is used to approximate an integral by the summation of the weighted function evaluated at quadrature points $(x_i, i = 1, \cdots, n)$: $\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i)$, where the weight $w_i$ and the quadrature points are based on the Hermite polynomial. GHQ can produce accurate approximations. However, as the dimensionality of the latent variables increases, say more than three latent variables, the estimation becomes infeasible. In addition, GHQ fails to find the maximum for certain functions (Huber et al., 2004). Adaptive Gauss-Hermite quadrature (AGHQ) is an improved method based on GHQ. Instead of using a fixed set of quadrature points and weights, AGHQ identifies the most relevant points and weights for each response pattern.

Although AGHQ reduces the required number of quadrature points and improves computational efficiency compared to GHQ, it is still time-consuming for high-dimensional models.

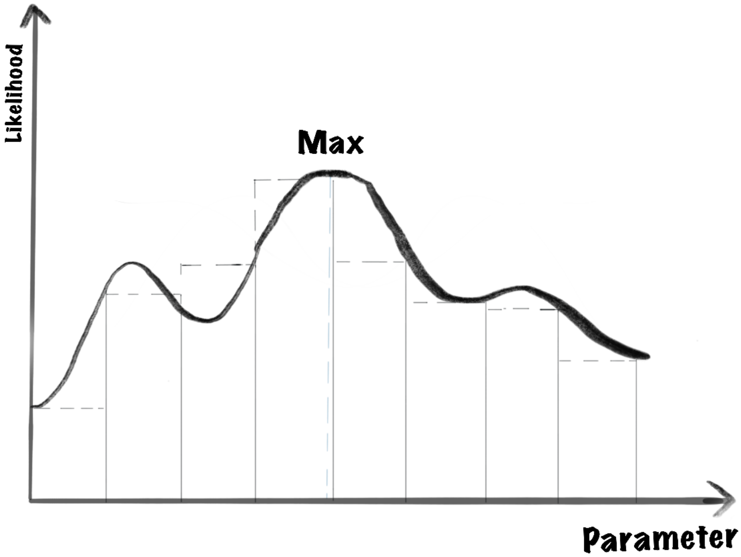


Figure 4.1: An example plot of computing the area under the likelihood curve. Quadrature-based methods divide the irregular curve into multiple small rectangles and sum the areas of the rectangles as the approximation of the area under the curve.

*2. Simulation-based approach.* Simulation-based approaches such as Monte Carlo expectation-maximization (Meng & Schilling, 1996), the Metropolis-Hastings Robbins-Monro methods (Cai, 2010), and Markov Chain Monte Carlo methods avoid directly computing the complicated integral in the likelihood function. Instead, this approach generates samples to approximate the desired integrals or distribution (i.e., a posterior distribution). These approaches can be implemented in various ways. In general, it is relatively feasible to deal with complex models. The computational complexity of the simulation-based approach is linear in the dimensionality of the latent variable models (Cai, 2010), while that of the quadrature-based approach is exponential. However, on the other hand, this approach is time-consuming in evaluating convergence and computing statistics like log-likelihood after the estimation procedure.

*3. Laplace approximations.* Another approach is to use Laplace approximations (Andersson et al., 2023; Huber et al., 2004; Niku et al., 2017) to approximate the integrals of the marginal log-likelihood, which is efficient and computationally simple (Shun, 1997). Laplace approximations are used to approximate the integral $\int e^{-Nh(\boldsymbol{x})}$ using basic numerical operations on the statistics from the sample and the model in addition to certain derivatives.

The Laplace approximations for GLLVMs are derived in Articles III and IV. The second-order Laplace approximation of the marginal log-likelihood function for individual $f$ can be expressed as (Shun, 1997):

$$\tilde{l}_f^{\text{Lap2}}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{P}{2}\log(2\pi) - \frac{1}{2}|\boldsymbol{H}_f| - \hat{h} + \log(1+\epsilon_f), \tag{4.3}$$

with

$$\epsilon_f = -\frac{1}{2}\left[\frac{1}{4}\sum_{jklm}^{P}\frac{\partial^4\hat{h}}{\partial z_j\partial z_k\partial z_l\partial z_m}b_{jl}b_{km} - \frac{1}{4}\sum_{jklrst}^{P}\frac{\partial^3\hat{h}}{\partial z_j\partial z_k\partial z_l}\frac{\partial^3\hat{h}}{\partial z_r\partial z_s\partial z_t}b_{jr}b_{kl}b_{st}\right. \tag{4.4}$$
$$\left. - \frac{1}{6}\sum_{jklrst}^{P}\frac{\partial^3\hat{h}}{\partial z_j\partial z_k\partial z_l}\frac{\partial^3\hat{h}}{\partial z_r\partial z_s\partial z_t}\frac{1}{6}b_{jr}b_{ks}b_{lt}\right].$$

In this expression, $h_f(\boldsymbol{z}) = -\log P(\boldsymbol{y}_f|\boldsymbol{z})\psi(\boldsymbol{z};\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\hat{h} = h_f(\hat{\boldsymbol{z}}_f)$, and $\boldsymbol{H}_f = \frac{\partial^2\hat{h}}{\partial\boldsymbol{z}\partial\boldsymbol{z}'}$. $\hat{\boldsymbol{z}}_f$ represents the posterior modes of the latent scores of individual $f \in 1,\cdots,N$. $b_{jk}$ represents the entry of row $j$ and column $k$ in $\boldsymbol{H}_f^{-1}$.

The first-order Laplace approximation can be obtained from Equation 4.3 by setting $\epsilon_f = 0$. Although first-order Laplace approximations are simpler and faster, second-order Laplace approximations provide a more accurate and robust approximation (Andersson et al., 2023; Andersson & Xin, 2021). Note that the $h_f$ function depends on the measurement model of the observed outcome variables. Hence, the implementation of Equation 4.3 will need to adapt to the distribution of the outcome variables $y_i$. The relevant derivatives of $h_f$ are derived analytically and programmed in C++ in this thesis.

So far, we have moved from the original marginal log-likelihood function (Equation 4.2) to the Laplace approximation function (Equation 4.3). The estimation problem is then to find estimates of item parameters that maximize Equation 4.3. To do so, we need to derive the gradient analytically. For $\theta \in \boldsymbol{\theta}$, its gradient can be expressed as:

$$\nabla_f^{\theta} = \frac{\partial l_f^{\text{Lap2}}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\theta} + \frac{\partial\hat{\boldsymbol{z}}_f}{\partial\theta}\frac{\partial l_f^{\text{Lap2}}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\boldsymbol{z}}\Bigg|^{\boldsymbol{z}=\hat{\boldsymbol{z}}_f}. \tag{4.5}$$

Note that the estimation can be simplified based on the model structure by filtering out zero and repeated entries (Andersson et al., 2023; Andersson & Xin, 2021). The optimization of Equation 4.3 is realized by a quasi-Newton method using the BFGS algorithm. The details of the iteration procedure of the algorithm can be found on Page 5 of Article III. Andersson and Jin (2022) implemented the proposed methods and AGHQ in an R package called *lamle.*

## 4.4 Ethical considerations

"Research ethics refer to a diversity of values, norms, and institutional arrangements that contribute to constituting and regulating scientific activity." –

the National Research Ethics Committee for the Social Sciences and Humanities (NESH).

For ethical considerations, this dissertation follows *Guidelines for Research Ethics in the Social Sciences and the Humanities* by the National Research Ethics Committee for the Social Sciences and Humanities in Norway (NESH, 2016). Some potential ethical issues regarding international large-scale assessments are discussed based on the guidelines of NESH.

First, we use data from PISA and PIAAC that cover human participants aged between 15 and 65 years. Before collecting personal data, researchers must obtain informed consent from the participants or their parents/guardians in the case of minors. It is important to note that the consent must be given freely, and it should be easy to understand, accessible, and clear. The participants have the right to decline or withdraw consent without any negative consequences. PISA and PIAAC obtain documented consent from parents/guardians of students and adult participants, respectively, prior to their participation in the surveys.

Second, the surveys collect participants' personal data, which must be processed and stored carefully. In PISA and PIAAC, the information of the participants is securely and confidentially treated and stored (OECD, 2014a, 2014d). In addition, personal details such as school names and student names are replaced with digital identifiers. We have no access to the coding scheme, so we cannot identify participants. That is, the data are anonymous to us. In this dissertation, the cognitive responses and log files are utilized, while the background information of the participants is only summarized on a descriptive level.

A third ethical concern is the impact of international large-scale assessments. For example, many researchers, educators, and policymakers are concerned about the results of PISA. The original aim of PISA is to monitor 15-year-old students' learning outcomes and figure out the factors that influence the results (Breakspear, 2012). However, the impacts of PISA go far beyond its original purpose. For example, policymakers seek to find the best educational practice based on the results of PISA and hope to apply that to their country (Breakspear, 2012). Nevertheless, whether "the best practice" is universal is questionable. Therefore, the use of the results of international large-scale assessments requires caution, and policymakers should be aware of context differences in the results. Considering this perspective, researchers should clearly describe methods and carefully interpret their results. This enables practitioners to evaluate if they can apply the results to their own practices. In this dissertation, we describe our methods in detail in each article. The recoding scheme and R code used in the empirical studies are accessible in the appendices of the articles or the Open Science Framework. Regarding the interpretations of the empirical results, we compare our results with existing research and connect the conclusions with theory.

Chapter 5

# Summary of the articles

## 5.1 Article I: Identify solution patterns

Zhang, M.*, & Andersson, B. (2023). Identifying problem-solving solution patterns using network analysis of operation sequences and response times. *Educational Assessment*, *28*, 172-189.

Article I investigates how respondents approach a solution to a computer-based task by making use of information from log files. Specifically, we aim to identify and visualize typical solution patterns of respondents. To achieve this, we first use network graphs to represent the transitions between actions and the time spent on the transitions. Next, we define seven network features to extract essential information from the response process using the process data of each respondent. These network features are determined to reflect the cognitive processes in problem-solving (Mayer & Wittrock, 2006): representing, planning/monitoring, executing, and self-regulating. After computing the network features for each respondent, we cluster the respondents based on the values of their network features through a clustering technique - GMMs. Respondents who share a similar profile on the network features are classified into the same group. Finally, we aggregate the process data of the respondents from each cluster and plot cluster-level network graphs of action sequences and response times. By combining these network graphs with the descriptive statistics of the seven network features for each cluster, we seek to understand and interpret the typical solution patterns.

To illustrate our approach, we use a PISA 2012 problem-solving task with a sample from the United States. Since successful and unsuccessful respondents may display distinct solution patterns, we apply the proposed method separately to the two groups. The results suggest two clusters for the failure group, and we interpret them as less-able and low-effort clusters. We also identify four clusters for the success group and interpret them as adaptable, back-and-forth, deliberate, and trial-and-error clusters. Students in the less-able cluster constantly tried various operations, but they got stuck in incorrect transitions. Similarly, students in the trial-and-error cluster randomly tried a variety of operations. Over time, however, they distinguished the correct operations from the incorrect ones, developing persistence in finding the correct solution. In contrast, students in the low-effort cluster tended to try an incorrect solution and gave up. While students from the adaptive cluster also chose incorrect solutions at first, they persisted in modifying their solutions to reduce the difference between the current and goal state. Students in the back-and-forth cluster tended to revisit their previous operations, which may indicate hesitance or self-regulating. Students from the deliberate cluster spent a considerable amount of time developing a strategy before putting it into action, indicating the cognitive process of planning. The results provide a more fine-grained understanding of the problem-solving processes that go beyond correctness/incorrectness and can potentially benefit educational practice.

## 5.2 Article II: Validate process indicators

Zhang, M.*, Andersson B., & Greiff S. (2023). Generalizing process data measures of planning and non-targeted exploration: Item-level and structural relationships in PIAAC. *Journal of Intelligence*, *11*, 156.

Representing and planning are two important cognitive processes in problem-solving (Mayer & Wittrock, 2006) and are closely associated with task performance. As these cognitive processes are difficult to observe directly, researchers have leveraged process data to gain insights into mental problem-solving activities. Accordingly, researchers have defined certain process indicators to infer these cognitive processes (Eichmann et al., 2019; Eichmann, Greiff, et al., 2020). However, the extent to which the process indicators can be generalized across different tasks has not been explicitly examined. Article II aims to validate two pre-defined process indicators for planning (Eichmann et al., 2019) and non-targeted exploration (Eichmann, Greiff, et al., 2020) across multiple dynamic tasks in the PIAAC 2012 PS-TRE domain.

The process indicators used in Article II are the longest duration (the planning indicator) and the number of initial non-targeted operations (the non-targeted exploration indicator). Since planning is resource-intensive and time-consuming, time-related measures have been suggested to reflect the planning process. In dynamic problems, the longest duration indicator has been proposed to capture the quantity of planning, which is defined as the longest time interval between two successive operations (Eichmann et al., 2019). As for the representing process, it is essential to collect information to get a better understanding of the nature of the problem and expand the problem space. To gather information in dynamic problems, respondents often need to actively engage with the computer. By examining the interactions between the computer and the respondent, we infer the extent to which an individual is exploring the problem environment. The human-computer interactions can be categorized into goal-directed and non-targeted operations depending on if they appear in any optimal solutions (Eichmann, Greiff, et al., 2020). Goal-direct operations contain similar information as task performance, while task scores have been included in the analysis. Non-targeted operations are the focus of our study because they provide additional information than task performance. Because only the initial non-targeted operations are related to generating new information and expanding the problem space, we use the number of initial non-targeted operations as the indicator for non-targeted exploration in this study.

Although the indicators for planning and non-targeted exploration have been used in previous studies (Eichmann, Goldhammer, Greiff, et al., 2020; Eichmann et al., 2019; Eichmann, Greiff, et al., 2020), the construct validity of the indicators has not been explicitly evaluated. In this article, our aim is twofold: a) to examine the construct validity of the indicators based on the dynamic problems in PIAAC 2012, and b) to test the relationships between planning, non-targeted exploration, and problem-solving competency. To accomplish the goals, we extract the indicators for planning and non-targeted exploration from the process data, recode the process indicators into equal-sized ordinal categories, and separately apply three uni-dimensional CFA models to the process indicators and task performance. Model fit indices and factor loadings are presented as evidence of the construct validity of the indicators. For the second aim of the study, we incorporate all three indicators into one model and estimate the covariance of the latent variables while considering the residual correlations among the indicators from the same item. The findings are summarized as follows. First, the results provide evidence of the construct validity of the planning indicators. Second, the non-targeted exploration indicator is less suitable to be analyzed simultaneously. Third, non-targeted exploration is strongly related to problem-solving competency in general, whereas planning and problem-solving competency are weakly negatively related. Fourth, these relationships vary substantially across tasks.

## 5.3 Article III: Laplace approximations of GLLVMs for categorical data

Andersson, B.*, Jin, S., & Zhang, M. (2023). Fast estimation of multiple group generalized linear latent variable models for categorical observed variables. *Computational Statistics & Data Analysis*, *182*, 107710.

Article III focuses on the issue of computational efficiency when estimating high-dimensional GLLVM with multiple groups for categorical observed data. Such models can be applied to analyze, for example, categorical response data from international large-scale assessments that evaluate respondents' abilities across different domains such as mathematics, reading, and science, across countries. The MML approach can be used for the estimation of GLLVMs. The marginal likelihood function consists of intractable integrals over the latent variables, requiring numerical approximations. Quadrature-based methods, such as GHQ and AGHQ, have commonly been used to approximate the integrals. They work well for models with one or two latent variables. However, when the dimension of the latent variables exceeds three, quadrature-based methods become infeasible since the computational difficulty increases exponentially with the dimension. A computationally efficient approach is to use Laplace approximations to approximate the integrals instead (Huber et al., 2004; Niku et al., 2017; Shun, 1997).

In Article III, we consider a second-order Laplace approximation for the integrals of marginal log-likelihood when estimating high-dimensional, multiple-group GLLVMs for categorical observed variables. Laplace approximations require high-order derivatives for each measurement model for the observed variables. The formula for the second-order Laplace approximation to the marginal log-likelihood is presented in Section 2.2 of Article III. We demonstrate an efficient way to implement this approach by considering the model structure and filtering any zero or repeated entries in the formula. We analytically derive the elements in the formula and implement the approach in an R package *lamle* (Andersson & Jin, 2022).

To assess the performance of the proposed method in terms of computational efficiency, convergence, and the recovery of item parameters, we conduct a simulation study including four conditions: 2 (the number of observed variables: three or four) × 2 (the type of model structure: independent-clusters model or a cross-loading model). We generate 1000 samples under each of the conditions with a sample size equal to 1000 and four latent variables. Second-order Laplace approximations are used for estimation and comparison against first-order Laplace approximations and AGHQ with three and five quadrature points. The results suggest that a) second-order Laplace approximations achieve 100% convergence rates. This is a significant improvement when comparing to first-order Laplace approximation, especially for cross-loading models with fewer items; b) second-order Laplace approximations achieve similarly accurate and precious estimates as AGHQ with five quadrature points, and both recover the item parameters better than first-order Laplace approximations and AGHQ with three quadrature points; and c) Laplace approximations take much shorter time than quadrature-based methods. In summary, second-order Laplace approximations produce fast (compared to AGHQ) yet accurate (compared to first-order Laplace approximations) parameter estimates for high-dimensional, multi-group GLLVMs for binary observed variables. In addition, we illustrate the method using empirical data from Hong Kong, Macao, Shanghai, and Chinese Taipei in the PISA 2009 assessment of mathematics, reading, and science. The empirical study involves four groups of respondents, 188 items in total, and the measurement of three latent abilities. First- and second-order Laplace approximations and AGHQ with three, five, and thirteen quadrature points are used for the estimation. The results closely resemble those of the simulation study. Namely, second-order Laplace approximations are computationally efficient and produce a similar log-likelihood value as AGHQ with 5 and 13 quadrature points, as indicated by the results.

## 5.4 Article IV: Laplace approximations of GLLVMs for mixed data

Zhang, M., Andersson, B.*, & Jin, S. Estimation of generalized linear latent variable models for performance and process data with ordinal, continuous, and count observed variables. *Submitted to British Journal of Mathematical and Statistical Psychology.*

A mixture of discrete and continuous data often occurs in data collection. For example, game-based assessments routinely record the number of correct/incorrect trials, the number of mouse clicks, time-on-screen, and performance scores. Similarly, computer-based assessments collect the complete human-computer interaction and provide information such as responses, response times on the task, response time until the first interaction, and the number of interactions. Such data consist of different types of variables: continuous, ordinal, and count data. Although the data offer valuable insights into response processes from various perspectives, the analysis of the data is challenging due to the complex dependencies of the observed variables and the inherent non-continuity and non-normality of the discrete variables. These make it difficult to directly apply conventional factor analysis (Jöreskog, 1969) or item factor analysis such as IRT models to the mixed types of observed variables. GLLVMs are promising to deal with this situation. GLLVMs can handle different types of outcome variables and link the expected value of outcome variables conditional on latent variables to a linear combination of predictors through a link function. However, the estimation of model parameters hinders the application of GLLVMs when the latent variables have a high dimensionality due to the intractable integrals of the marginal log-likelihood.

In Article IV, we use first- and second-order Laplace approximations to estimate GLLVMs for a combination of ordinal, continuous, and count observed variables. Similar to Article III, we analytically derive the derivatives related to the measurement models with normal, Poisson, and negative-binomial distributed data and program the derivatives in C++. Two simulation studies are performed to evaluate the performance of first- and second-order Laplace approximations, and an empirical study is conducted to illustrate the approach and provide references to the simulation design. Specifically, we use a PISA 2012 computer-based mathematics item that consists of three subtasks. Categorical response scores, the time spent on the task, and the number of actions for 1029 Australian respondents on each subtask are utilized as the observed indicators. Three uni-dimensional measurement models are separately applied to the three types of indicators, and the estimates provide a foundation for the range of true values of the item parameters in the simulation studies. We also place the three types of indicators into a single model using GLLVMs to demonstrate the application of the proposed method. We then conduct simulation studies to examine the estimation efficiency and parameter recovery of Laplace approximations in GLLVMs for a mixture of continuous (X), ordinal (Y), and count (Z) observed variables. In Simulation 1, we consider 2 (the distribution of the count data model: Poisson or negative-binomial distributions) × 2 (the number of items: three or six) × 2 (the covariance between the latent variables: small or large) = 8 conditions. In Simulation 2, we further consider the dependency of the indicators from the same item or sub-task by adding a residual factor for each item. For example, we add a latent residual factor R1 to account for the residual correlations between the indicators X1, Y1, and Z1 from the first item. For the purposes of simplicity, we assume equal residual factor loadings across the indicators from the same item. The magnitude of residual factor loadings (small or large) is then added to Simulation 2, resulting in 16 conditions. Both simulations conduct 1000 replications under each condition with a sample size equal to 1000. The results suggest that second-order Laplace approximations achieve a higher average convergence rate and produce more accurate estimates for the model parameters. However, they take a longer time for estimation compared to first-order Laplace approximations, especially when the model is complex (i.e., models involving residual factors).

Chapter 6

# Discussion and implication

This dissertation aims to use process data to better understand problem-solving theoretically (Articles I and II) and methodologically (Articles III and IV). Four articles are included in the dissertation.

Articles I and II focus on the cognitive processes in problem-solving through the information present in respondents' action sequences and response times. In Article I, we propose a method for identifying respondents' solution patterns based on the network features extracted from their process data. The definition and extraction of network features are guided by a theoretical framework (Mayer & Wittrock, 2006) to reflect the cognitive processes in problem-solving, making the results easier to interpret compared to the previous studies (Vista et al., 2016; Vista et al., 2017; Zhu et al., 2016). In Article II, we examine the construct validity of the process indicators for planning and non-targeted exploration using the PIAAC PS-TRE problems. Additionally, we test the overall and task-specific relationships between planning, non-targeted exploration, and problem-solving competency.

Article III introduces a computationally efficient method for applying higher-order Laplace approximations to the integrals of the marginal log-likelihood of GLLVMs for categorical observed variables. Article IV further applies the method proposed in Article III to GLLVMs for a mixture of ordinal, continuous, and count observed variables. This allows for simultaneous analysis of task performance, time-on-task, and the number of actions. Namely, GLLVMs enable the joint modeling of performance data and process indicators, as well as other scenarios with a mixture of discrete and continuous variables. However, the estimation of high-dimensional GLLVMs is computationally demanding and the available implementation tools are limited, which hinders the application of GLLVMs. Our approach seeks to address the computational issue and hopes to make GLLVMs more accessible for practitioners.

In this chapter, I review how our studies cope with the challenges in process data analysis and discuss the contributions and limitations of the project.

## 6.1 Dealing with challenges of process data analysis

Section 2.2.2 discusses six challenges of analyzing process data, and here I explain how our studies cope with these challenges. A summary is provided in Table 6.1. The first challenge is the large volume of process data consisting of a great variety of variables. To address this challenge, we subset the data according to specific countries or tasks and use certain methods to recode the data in the articles. The data preprocessing procedures are clearly described in the articles. Second, process data usually have high dimensions. In Article I, we have seven network features and apply cluster analysis that can handle the challenge of high dimensions. In Article II, we have a total of 21 observed indicators and analyze them in the three-dimensional latent variable model. In Articles III and IV, we propose using Laplace approximations to increase computational efficiency in GLLVMs. This can also be applied to other situations that involve a mixture of continuous and discrete observed variables. Third, individual process data vary in the lengths of action sequences and response times. In our articles, we extract a fixed number of features from the process data for each individual rather than working on the original single actions.

Table 6.1: Dealing with challenges of process data analysis.

| Challenges | Solutions |
|---|---|
| Large volume & great variety | Articles I-IV: Subset and recode the data |
| High dimension | Article I: Use a machine learning approach (cluster analysis)<br>Article II: Incorporate 21 indicators into a three-dimensional model<br>Articles III & IV: Use Laplace approximations to increase estimation efficiency |
| Varied lengths | Articles I, II, and IV: Extract a fixed number of features from the process data for each individual |
| Data dependencies | Article I: Use network features such as reciprocity to reflect the dependency<br>Article II: Consider residual correlations of the indicators from the same item<br>Article IV: Consider residual factors |
| Noise | Article I-IV: Recode the data and data cleaning (e.g., remove outliers) |
| Interpretation & validation | Article I: Use a theory to guide the extraction of network features<br>Article II: Examine the construct validity of the pre-defined process indicators |

In addition, it is important to consider the dependencies of process data. In Article I, we defined edge density, reciprocity, transitivity, density, and the External-Internal index to reflect the dependencies of the actions. In Articles II and IV, we add residual correlations or residual factors to account for the correlations among the indicators defined for the same item. Fifth, the noise in the process data is handled through data recoding and data cleaning. For example, we define a category of insignificant operations to cover the operations beyond the task instructions in Article I. In Articles II and IV, we exclude respondents with extreme values on the process indicators. Last, interpretation and validation are challenging when analyzing process data. Unlike a purely data-driven approach, our feature extraction is based on the problem-solving theory by Mayer and Wittrock (2006) in Articles I and II, enhancing the interpretability of our findings. Furthermore, we particularly examine the construct validity of the process indicators for planning and non-targeted exploration defined in

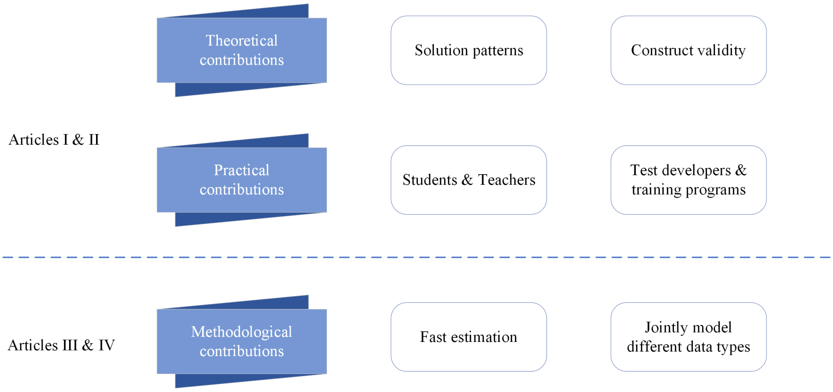


Figure 6.1: Contributions of the dissertation.

previous studies (Eichmann, Goldhammer, Greiff, et al., 2020; Eichmann et al., 2019; Eichmann, Greiff, et al., 2020) in Article II.

In summary, our work provides insight into how to address the challenges in terms of high dimensions, data dependencies, and validation. Some contributions and limitations of this dissertation are further discussed in the following subsections.

## 6.2 Contributions

Articles I and II shed light on the problem-solving processes using process data (theoretical contributions) and provide implications for educational practice that can positively impact students, teachers, and test developers (practical contributions). By comparison, Articles III and IV contribute primarily from a methodological perspective by improving the computational efficiency of GLLVMs using Laplace approximations (methodological contributions). In this section, I discuss the theoretical, practical, and methodological contributions of the dissertation in detail. See Figure 6.1 for an overview.

### 6.2.1 Theoretical contributions

The competency of problem-solving is essential for individuals when adapting to a rapidly changing world and is required in personal and professional contexts. Both formal education and adult education should help individuals become better problem-solvers. To achieve this goal, it is crucial to gain a better understanding of how respondents solve problems. Our first two articles illuminate the cognitive processes involved in problem-solving, deepen the understanding of respondents' solution patterns, and provide evidence for the validity of the process indicators that reflect the planning and non-targeted exploration processes.

In Article I, we identify and visualize solution patterns for the respondents based on their process data. These solution patterns distinguish the various strategies used in the task and indicate individual differences. With an example task from PISA 2012, we identify four solution patterns (i.e., adaptable, back-and-forth, deliberate, and trial-and-error) used by respondents who successfully solved the task and two patterns (i.e., less-able and low-effort) by respondents who failed to solve the task. The characteristics and interpretations of each pattern are described in detail in our paper, enhancing the understanding of how successful and unsuccessful problem-solvers approach the task. Compared to prior studies that profile students based on other behavioral indicators, such as the use of the VOTAT strategy (Gao et al., 2022; Gnaldi et al., 2020; Greiff et al., 2018; Stadler, Niepel, et al., 2019), our study uses network features that capture more comprehensive information from both action sequences and response times, and no predefined strategies are required to apply this method. In addition, the natural advantage of using network analysis is that it provides a direct visualization of the entire human-computer interaction. Both individual-level and cluster-level networks of process data have been presented in Article I, providing a straightforward illustration of the solution patterns.

In Article II, we provide evidence for the validity of the process indicators for planning and non-targeted exploration using the PIAAC 2012 PS-TRE tasks. Our results suggest evidence of internal construct validity of the planning indicator, but weaker evidence for the non-targeted exploration indicator. That is, the latent variable planning can capture a great portion of the shared variance among the observed indicator - the longest duration, whereas the latent variable non-targeted exploration shows varied impacts on the indicators (i.e., the number of non-targeted operations) across tasks. We further dig into the characteristics of the tasks to seek potential explanations for the results. In addition, we test the relationships between planning, non-targeted exploration, and problem-solving competency and examine whether the relationships are consistent or varied across tasks. Our results indicate task-specific relationships, suggesting the importance of considering the residual correlations between indicators from the same item. This enhances the understanding of the validity of using the process indicators defined by previous studies (Eichmann, Goldhammer, Greiff, et al., 2020; Eichmann et al., 2019; Eichmann, Greiff, et al., 2020) and of the functions of planning and non-targeted exploration in various tasks.

In summary, Articles I and II are grounded in the problem-solving theory of Mayer and Wittrock (2006) and use the process data to improve the understanding of the solution patterns and cognitive processes involved in problem-solving. The studies provide empirical evidence for the validity of indicators of cognitive problem-solving processes, test hypotheses about the relationships between cognitive processes and task performance, and discover behavioral patterns in problem-solving. The findings help researchers gain deeper insights into the theoretical concepts of problem-solving.

### 6.2.2 Practical contributions

In addition to theoretical contributions to the study of problem-solving, the empirical findings from Articles I and II can further contribute to educational practice. This can be viewed from four perspectives: students, teachers, test developers, and problem-solving training programs.

The first practical contribution is directed toward students. For example, Article I provides network graphs of process data for individuals. Students can review their own network graphs of process data to reflect on the cognitive processes. It is similar to a tape recording of respondents' response processes, but a tape recording requires more time and resources to review and analyze it. Reviewing individual network graphs of process data can help students engage in introspection to increase self-awareness, self-observation, self-monitoring, and self-reflection, which can further improve problem-solving skills (Jäkel & Schreiber, 2013).

Second, our results can provide educators with materials to tailor their instructions according to students' solution patterns. For example, teachers can assign easier tasks or provide explicit hints to assist less-able students in finding solutions and improving their problem-solving skills. For low-effort students, teachers can talk to them and inquire about the cause of their de-motivation (e.g., inadequate ICT skills). Teachers can also adapt their instructions for different tasks. For example, for complex tasks with a substantial amount of information, teachers can guide students to explore the problem environment initially and foster the development of a plan.

Third, our approach can help test developers validate the task design. For instance, they can compare the solution patterns identified from the process data to the desired solution patterns. An unexpected deviation, such as many students checking the help button, may indicate that the task instructions are not clear enough. As another example, if a task is intended to measure the planning aspect of problem-solving, then there should be a positive relationship between planning and task performance. If the result shows a different direction, the task design should be reconsidered.

Fourth, getting a better understanding of problem-solving processes can contribute to enhancing problem-solving training programs. Consider incorporating a computer-simulated agent into a training program to provide real-time instructions and feedback to aid problem-solving. For example, if an individual has spent a long time planning in a dynamic problem but has not made any interactions, the agent can offer a hint to encourage exploratory behaviors if the individual is unfamiliar with the task environment. As another example, if the computer agent identifies an individual's solution pattern, it can select the next task to facilitate targeted development in weak areas. For instance, if a student is not good at making plans, a training program could provide additional tasks that require planning processes and give explicit hints about planning strategies, such as decomposing the problem. Such programs can potentially be implemented via a virtual learning environment. A virtual learning environment provides students with a designed information-rich and social space in which students can engage in a variety of learning activities (Dillenbourg et al., 2002). Previous studies have developed a problem-based learning model through a virtual learning environment, and the results suggest an increase in learning interest and satisfaction (Phungsuk et al., 2017). Similarly, our study provides materials for such programs, but the implementation requires more effort from the expert group and technicians.

### 6.2.3 Methodological contributions

Compared to the theoretical and practical contributions of Articles I and II, Articles III and IV mainly contribute from a methodological perspective. In the last two articles, we provide useful estimation tools for researchers and practitioners in estimating GLLVMs for various types of observed variables.

A mixture of discrete and continuous data is prevalent in various fields. In Articles II and IV, we illustrate this situation in computer-based assessments that include ordinal response scores, continuous response times, and counts of interactive behaviors. A joint model of these variables poses estimation difficulties, especially in high-dimensional models. We employ GLLVMs to analyze different types of observed variables simultaneously and use higher-order Laplace approximations to increase the estimation efficiency. The simulation studies indicate a significant reduction in the estimation time compared to quadrature-based methods and an improvement in terms of convergence rates and parameter recovery compared to first-order Laplace approximations.

Our method can be applied to other fields where a combination of different data types occurs. For example, it is possible to apply the method to ecological data and patient data. Ecological data often consist of species counts and biomass (Niku et al., 2017), and patient data typically include the presence/absence, frequency, and severity of certain symptoms (Daniels & Normand, 2006). The algorithms have been implemented in an R package called *lamle* (Andersson & Jin, 2022), which will soon be available on CRAN. Researchers and analysts can then directly use the package to deal with GLLVMs for continuous, ordinal, and count data simultaneously, instead of proceeding with one data type at one time as the *gllvm* package requires (Niku et al., 2017).

## 6.3 Limitations and future studies

Some limitations of the dissertation and implications for future studies should be noted.

First, we extract a limited number of process indicators from the extensive process data. For example, we extract the longest duration and the number of non-targeted operations in Article II, which capture only limited information of interest from the original, complete human-computer interaction. Much information has been ignored in the feature extraction procedure. However, it is possible to include additional process indicators to reflect other aspects of the response process. For example, other network features such as the centrality of each operation and other process indicators reflecting the self-regulating process can be included in future studies to verify the interpretations of the analytical results.

Second, the validation of the empirical findings requires further investigation. Specifically, the solution patterns identified in Article I are determined through a model comparison approach and are interpreted in an ad-hoc manner. Whether the respondents adopt the solution patterns as we interpret them in the article cannot be verified because we use secondary data and no relevant information on solution patterns is available in the dataset. If it is possible to collect data in the future, it would be beneficial to incorporate qualitative research methods, such as interviews and think-aloud protocols or questionnaires about the problem-solving strategies used in the task.

Third, the applicability of the results of the study to different situations and settings should be further explored in future studies. Articles I, II, and IV consider only a single set of data from international large-scale assessments to illustrate our approach. However, the extent to which the findings from one country can be generalized to other countries needs careful consideration. As an example, we identified six solution patterns using the U.S. data in Article I, but the results using datasets from other countries, such as China or Japan, may be different due to cultural differences. In a similar vein, the relationships between planning, non-targeted exploration, and problem-solving competency in Article II may vary depending on the type of problem. Therefore, examining the generalizability of the empirical findings across countries or tasks is an interesting direction for future studies.

Fourth, in Articles III and IV, the Laplace approximations need complex analytical derivatives for specific distributions. For other distributions not covered in the articles, up to fifth-order derivatives must be analytically derived. That is, if researchers assume other distributions of the observed variables, extra work will be required. More distributions are considered to be included and implemented in the package *lamle*.

## 6.4 Concluding remarks

This dissertation analyzes the process data from international large-scale assessments to understand how problem-solvers approach problems and to reflect the cognitive processes underlying the observed response processes. In addition, we propose to use a fast estimation method (i.e., higher-order Laplace approximations) to jointly analyze performance data and process data within the framework of GLLVMs.

Process data provide researchers with valuable information about respondents' response processes and can serve as a window into their minds as they solve problems. Our studies demonstrate the potential of using process data to enhance the understanding of cognitive problem-solving processes. In addition, we examine evidence for the validity of analyzing process indicators across tasks, suggesting that it is important to examine the construct validity of process indicators and to consider the dependency of process indicators derived from the same task.

A joint model of performance data and process data requires efficient estimation algorithms. The use of higher-order Laplace approximations can significantly increase the estimation speed, and the method is applicable to a combination of continuous, ordinal, and count data. The method has been implemented in the *lamle* package, facilitating its convenience for practitioners.

# Bibliography

Aburezeq, K., & Kasik, L. (2021). The relationship between social problem solving and psychological well-being: A literature review. *Romanian Journal of Psychological Studies*, *9*, 3–16.

AERA. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Albert, D., & Steinberg, L. (2011). Age differences in strategic planning as indexed by the Tower of London. *Child Development*, *82*(5), 1501–1517.

Andersson, B., & Jin, S. (2022). Lamle: Maximum likelihood estimation of latent variable models using adaptive quadrature and laplace approximations.

Andersson, B., Jin, S., & Zhang, M. (2023). Fast estimation of multiple group generalized linear latent variable models for categorical observed variables. *Computational Statistics & Data Analysis*, *182*, 107710.

Andersson, B., & Xin, T. (2021). Estimation of latent regression item response theory models using a second-order laplace approximation. *Journal of Educational and Behavioral Statistics*, *46*(2), 244–265.

Apuke, O. D. (2017). Quantitative research methods: A synopsis approach. *Kuwait Chapter of Arabian Journal of Business and Management Review*, *33*(5471), 1–8.

Arieli-Attali, M., Ou, L., & Simmering, V. R. (2019). Understanding test takers' choices in a self-adapted test: A hidden Markov modeling of process data. *Frontiers in Psychology*, *10*, 83.

Arlin, P. K. (1989). The problem of the problem. In J. D. Sinnott (Ed.), *Everyday problem solving: Theory and applications* (1st ed.). Westport: Praeger Publishers.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, *118*(4), 1279–1333.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach.* John Wiley & Sons.

Bell, B. S., & Kozlowski, S. W. J. (2008). Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology*, *93*(2), 296.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459.

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1), e1230.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13–38.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71.

Bransford, J. D., & Stein, B. S. (1984). *The IDEAL problem-solver: Improving learning, thinking, and creativity.* Freeman.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. OECD Publishing.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Macmillan.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Casella, G., & Berger, R. L. (2021). *Statistical inference* (2nd ed.). Cengage Learning.

Cassidy, T., & Long, C. (1996). Problem-solving style, stress and psychological illness: Development of a multifactorial measure. *British Journal of Clinical Psychology*, *35*(2), 265–277.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *Psychometrika*, *85*(4), 1052–1075.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, *10*, 486.

Chung, G. K., De Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior*, *18*(6), 669–684.

Danek, A. H., Wiley, J., & Öllinger, M. (2016). Solving classical insight problems without aha! experience: 9 dot, 8 coin, and matchstick arithmetic problems. *The Journal of Problem Solving*, *9*(1), 4.

Daniels, M. J., & Normand, S.-L. T. (2006). Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*, *7*(1), 1–15.

De Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology*, *10*, 1280.

Dillenbourg, P., Schneider, D., & Synteta, P. (2002). Virtual learning environments. *Proceedings of the 3rd Hellenic Conference Information & Communication Technologies in Education*, *2002*, 01.

Dormann, T., & Frese, M. (1994). Error training: Replication and the function of exploratory behavior. *International Journal of Human-Computer Interaction*, *6*(4), 365–372.

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020). Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology*, *112*(8), 1546.

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, *128*, 1–12.

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning*, *36*(6), 933–956.

Fischer, A., Greiff, S., & Funke, J. (2011). The process of solving complex problems. *Journal of Problem Solving*, *4*(1), 19–42.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, *7*(1), 69–89.

Gallagher, S. A., Stepien, W. J., & Rosenthal, H. (1992). The effects of problem-based learning on problem solving. *Gifted Child Quarterly*, *36*(4), 195–200.

Gao, Y., Zhai, X., Bulut, O., Cui, Y., & Sun, X. (2022). Examining humans' problem-solving styles in technology-rich environments using log file data. *Journal of Intelligence*, *10*(3), 38.

Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist*, *21*(1-2), 99–120.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38.

Gnaldi, M., Bacci, S., Kunze, T., & Greiff, S. (2020). Students' complex problem solving profiles. *Psychometrika*, 1–33.

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, *126*, 248–263.

Greiff, S., Scheiter, K., Scherer, R., Borgonovi, F., Britt, A., Graesser, A., Kitajima, M., & Rouet, J.-F. (2017). Adaptive problem solving: Moving towards a new assessment domain in the second cycle of PIAAC.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? a showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105.

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, (13), 74–83.

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development*, *61*, 407–421.

Han, Y., Liu, H., & Ji, F. (2022). A sequential response model for analyzing process data on technology-based problem-solving tasks. *Multivariate Behavioral Research*, *57*(6), 960–977.

Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology*, *10*.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining*, *7*(1), 33–50.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, *166*, 104170.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In *Handbook of research on technology tools for real-world skill development* (pp. 750–777). IGI Global.

Huber, P., Ronchetti, E., & Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(4), 893–908.

Jäkel, F., & Schreiber, C. (2013). Introspection in problem solving. *The Journal of Problem Solving*, *6*(1), 20–33.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Johnston, M. P. (2014). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries*, *3*(3), 619–626.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, *48*(4), 63–85.

Jonassen, D. H. (2010). *Learning to solve problems: A handbook for designing problem-solving learning environments*. Routledge.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*(3), 381–389.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*(3), 347–387.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258.

Klahr, D. (2000). Scientific discovery as problem solving. In D. Klahr (Ed.), *Exploring science: The cognition and development of discovery processes.* MIT Press.

Klieme, E. (2000). Assessment of cross-disciplinary problem solving competencies. OECD Publishing.

Lesh, R., & Judith, Z. (2007). Problem solving and modeling. In K. F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning.* Information Age Publishing.

Levy, R. (2019). Dynamic bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, *54*(6), 771–794.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936–949.

Li, S., Pöysä-Tarhonen, J., & Häkkinen, P. (2022). Patterns of action transitions in online collaborative problem solving: A network analysis approach. *International Journal of Computer-Supported Collaborative Learning*, *17*(2), 191–223.

Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: An investigation of united states adults' employment status in PIAAC. *Frontiers in Psychology*, *10*, 646.

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified Multilevel Mixture IRT model. *Frontiers in Psychology*, *9*, 1372.

Liu, Y., Cheng, Y., & Liu, H. (2020). Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement*, *80*(4), 775–807.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Lord, F. M. (1986). Maximum likelihood and bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 157–162.

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action–effective strategic behaviors while solving complex problems. *Intelligence*, *64*, 98–112.

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2022). G's little helpers – VOTAT and NOTAT mediate the relation between intelligence and complex problem solving. *Intelligence*, *95*, 101685.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 176–201.

Mandler, J. M., & Mandler, G. (1964). Thinking: From association to gestalt.

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science*, *26*(1-2), 49–63.

Mayer, R. E. (1999). Problem solving. In S. R. P. Mark A. Runco (Ed.), *Encyclopedia of creativity* (pp. 437–447). Academic Press.

Mayer, R. E. (2014). What problem solvers know: Cognitive readiness for adaptive problem solving. In *Teaching and measuring cognitive readiness* (pp. 149–160). Springer.

Mayer, R. E. (2019). Problem solving. In *Oxford research encyclopedia of education.*

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. *Handbook of Educational Psychology*, 47–62.

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287–303). Routledge.

Meng, X.-L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*(435), 1254–1267.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11*(1), 3–31.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*, 391–411.

Mumford, M. D., Schultz, R. A., & Van Doorn, J. R. (2001). Performance in planning: Processes, requirements, and errors. *Review of General Psychology*, *5*(3), 213–240.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide: Statistical analysis with latent variables*. Muthén & Muthén.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

NESH. (2016). *Guidelines for research ethics in the social sciences and the humanities*. https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/guidelines-for-research-ethics-in-the-social-sciences-humanities-law-and-theology/

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem-solving program. *IFIP congress*, *256*, 64.

Newell, A., & Simon, H. (1956). The logic theory machine – a complex information processing system. *IRE Transactions on Information Theory*, *2*(3), 61–79.

Newman, M. (2018). *Networks*. Oxford University Press.

Niku, J., Warton, D. I., Hui, F. K., & Taskinen, S. (2017). Generalized linear latent variable models for multivariate count and biomass data in ecology. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(4), 498–522.

OECD. (2003). *The PISA 2003 assessment framework - mathematics, reading, science and problem solving knowledge and skills*. OECD Publishing.

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. OECD Publishing.

OECD. (2014a). *PIAAC technical standards and guidelines*. OECD Publishing.

OECD. (2014b). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems*. OECD Publishing.

OECD. (2014c). *PISA 2012 technical report*. OECD Publishing.

OECD. (2014d). *Technical standards for PISA 2012*. OECD Publishing.

OECD. (2017). *PISA 2015 results (volume v): Collaborative problem solving*. OECD Publishing.

OECD. (2019). Technical report of the survey of adult skills (PIAAC) (3rd ed.). OECD Publishing.

OECD. (2021). *The assessment frameworks for Cycle 2 of the Programme for the International Assessment of Adult Competencies*. OECD Publishing.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. *Advances in the Psychology of Thinking*, *1*, 1–44.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, *47*, 337–347.

Ouyang, F., Xu, W., & Cukurova, M. (2023). An artificial intelligence-driven learning analytics method to examine the collaborative problem-solving process from the complex adaptive systems perspective. *International Journal of Computer-Supported Collaborative Learning*, *18*(1), 39–66.

Phungsuk, R., Viriyavejakul, C., & Ratanaolarn, T. (2017). Development of a problem-based learning model via a virtual learning environment. *Kasetsart Journal of Social Sciences*, *38*(3), 297–306.

Polya, G. (2004). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*, *9*, 2231.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190.

Rehm, M., Rohlfing, K., & Goecke, K. U. (2003). Situatedness: The interplay between context(s) and situation. *Journal of Cognition and Culture*, *3*(2), 132–156.

Reichenberg, R. (2018). Dynamic bayesian networks in educational measurement: Reviewing and advancing the state of the field. *Applied Measurement in Education*, *31*(4), 335–350.

Ren, Y., Luo, F., Ren, P., Bai, D., Li, X., & Liu, H. (2019). Exploring multiple goals balancing in complex problem solving based on log data. *Frontiers in Psychology*, *10*, 1975.

Ristad, E. S., & Yianilos, P. N. (1998). Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(5), 522–532.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

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

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometric Sosiety.

Schoenfeld, A. H. (2016). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (reprint). *Journal of Education*, *196*(2), 1–38.

Segers, M. S. (1997). An alternative for assessing problem-solving skills: The overall test. *Studies in Educational Evaluation*, *23*(4), 373–98.

Shun, Z. (1997). Another look at the salamander mating data: A modified laplace approximation approach. *Journal of the American Statistical Association*, *92*(437), 341–349.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117.

Simon, H. A. (1979). Information processing models of cognition. *Annual Review of Psychology*, *30*(1), 363–396.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In G. L. W. (Ed.), *Knowledge and cognition* (pp. 105–127). Hillsdale, NJ: Erlbaum.

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26*(2), 145–159.

Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*, *10*, 777.

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. *Intelligence*, *72*, 1–12.

Stoeffler, K., Rosen, Y., Bolsinova, M., & von Davier, A. A. (2020). Gamified performance assessment of collaborative problem solving skills. *Computers in Human Behavior*, *104*, 106036.

Suryanto, H., Degeng, I. N. S., Djatmika, E. T., & Kuswandi, D. (2021). The effect of creative problem solving with the intervention social skills on the performance of creative tasks. *Creativity Studies*, *14*(2), 323–335.

Swaak, J., & de Jong, T. (1996). Measuring intuitive knowledge in science: The development of the what-if test. *Studies in Educational Evaluation*, *22*(4), 341–62.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, 1–20.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2020). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, (1), 1–33.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Transaction Publishers.

Tóth, K., Rölke, H., Goldhammer, F., & Barkow, I. (2017). Educational process mining: New possibilities for understanding students' problem-solving skills. In B. Csapó & J. Funke (Eds.), *The nature of problem solving*. OECD Publishing. OECD Publishing.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 1–10.

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika*, *86*, 190–214.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*, 83–112.

Unterrainer, J. M., Rahm, B., Leonhart, R., Ruff, C., & Halsband, U. (2003). The Tower of London: The impact of instructions, cueing, and learning on planning abilities. *Cognitive Brain Research*, *17*(3), 675–683.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Veerasamy, A. K., D'Souza, D., Lindén, R., & Laakso, M.-J. (2019). Relationship between perceived problem-solving skills and academic performance of novice learners in introductory programming courses. *Journal of Computer Assisted Learning*, *35*(2), 246–255.

Vista, A., Awwal, N., & Care, E. (2016). Sequential actions as markers of behavioural and cognitive processes: Extracting empirical pathways from data streams of complex tasks. *Computers & Education*, *92*, 15–36.

Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Computers in Human Behavior*, *76*, 656–671.

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–705.

Vörös, Z., Kehl, D., & Rouet, J.-F. (2021). Task characteristics as source of difficulty and moderators of the effect of time-on-task in digital problem-solving. *Journal of Educational Computing Research*, *58*(8), 1494–1514.

Weisberg, R. W. (2015). Toward an integrated theory of insight in problem solving. *Thinking & Reasoning*, *21*(1), 5–39.

Wenke, D., Frensch, P. A., & Funke, J. (2005). Complex problem solving and intelligence: Empirical relation and causal direction. In S. R. J. & P. J. E. (Eds.), *Cognition and intelligence: Identifying the mechanisms of the mind* (1st ed., pp. 160–187). Cambridge University Press.

Wertheimer, M., & Wertheimer, M. (1959). *Productive thinking.* Harper.

Wingersky, M. S. (1983). Logist: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Educational Research Institute of British Columbia Vancouver.

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79.

Wolcott, M. D., & Lobczowski, N. G. (2021). Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning*, *13*(2), 181–188.

Wood, P. K. (1983). Inquiring systems and problem structure: Implications for cognitive development. *Human Development*, *26*(5), 249–265.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving — more than reasoning? *Intelligence*, *40*(1), 1–14.

Xiao, Y., He, Q., Veldkamp, B., & Liu, H. (2021). Exploring latent states of problem-solving competence using hidden Markov model on process data. *Journal of Computer Assisted Learning*, *37*(5), 1232–1247.

Xiao, Y., & Liu, H. (2023). A state response measurement model for problem-solving process data. *Behavior Research Methods*, 1–20.

Xu, H., Fang, G., & Ying, Z. (2020). A latent topic model with Markov transition for process data. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 474–505.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(3), 392–423.

Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of collaborative problem solving based on process stream data: A new paradigm for extracting indicators and modeling dyad data. *Frontiers in Psychology*, *10*, 369.

Zhan, P., Chen, Q., Wang, S., & Zhang, X. (2023). Longitudinal joint modeling for assessing parallel interactive development of latent ability and processing speed using responses and response times. *Behavior Research Methods*, 1–22.

Zhan, P., & Qiao, X. (2022). Diagnostic classification analysis of problem-solving competence using process data: An item expansion method. *Psychometrika*, *87*(4), 1529–1547.

Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement*, *53*(2), 190–211.

Zoanetti, N., & Griffin, P. (2017). Log-file data as indicators for problem-solving processes. In B. Csapó & J. Funke (Eds.), *The nature of problem solving.* OECD Publishing. OECD Publishing.

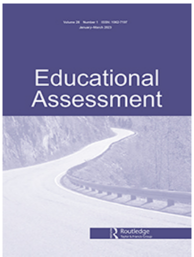
# Papers

Paper I

# Identifying Problem-solving solution patterns using network analysis of operation sequences and response times

**Maoxin Zhang, Björn Andersson**

Published in *Educational Measurement*.



Educational Assessment

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/heda20

# Identifying Problem-Solving Solution Patterns Using Network Analysis of Operation Sequences and Response Times

Maoxin Zhang & Björn Andersson

**To cite this article:** Maoxin Zhang & Björn Andersson (2023): Identifying Problem-Solving Solution Patterns Using Network Analysis of Operation Sequences and Response Times, Educational Assessment, DOI: 10.1080/10627197.2023.2222585

**To link to this article:** https://doi.org/10.1080/10627197.2023.2222585

View supplementary material

Published online: 16 Jun 2023.

Submit your article to this journal

Article views: 168

View related articles

View Crossmark data

EDUCATIONAL ASSESSMENT
https://doi.org/10.1080/10627197.2023.2222585

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

# Identifying Problem-Solving Solution Patterns Using Network Analysis of Operation Sequences and Response Times

Maoxin Zhang and Björn Andersson

University of Oslo, Oslo, Norway

**ABSTRACT**
Process data from educational assessments enhance the understanding of how students answer cognitive items. However, effectively making use of these data is challenging. We propose an approach to identify solution patterns from operation sequences and response times by generating networks from process data and defining network features that extract essential information from them. With these features, we group respondents to a problem-solving task from PISA 2012 using Gaussian mixture models. The results indicate the presence of two and four clusters for groups defined by failure and success on the task, respectively. We interpret the clusters as less-able, low-effort, adaptable, back-and-forth, deliberate, and trial-and-error clusters by considering the cluster-specific feature statistics. The proposed approach sheds light on students' problem-solving mental processes, which can aid item development and facilitate individualized feedback to students. The method is applicable to many computer-based problems, but a limitation is that the feature definitions can be task-dependent.

## Introduction

Problems are everywhere, no matter in daily life or at the workplace. Thus, problem-solving competency is highly demanded in modern society (OECD, 2014a) and has gained increasing attention in large-scale computer-based assessments. In such assessments, not only *how well* (the final performances) but also *how* (problem-solving processes) participants solve a problem can be recorded in log files and converted to process data. This study focuses on the process data of problem-solving tasks and proposes an approach that combines network analysis and Gaussian mixture models to investigate test-takers' solution patterns. Exploring process data helps researchers and educators understand the mental processes of students when solving a problem and can aid educational practices. In the remainder of this section, we review the problem-solving literature and then summarize the characteristics and analysis methods of process data. Subsequently, we introduce network analysis and the details of the present study.

### *Problem solving*

When encountering a problem without an obvious solution, problem-solvers need to engage in the cognitive processes of problem solving directed toward a goal (Mayer & Wittrock, 2006). Problem solving has four primary characteristics: (a) it is cognitive, making it hard to measure directly but possible to infer from the behavior of problem-solvers; (b) it is guided by certain goals; (c) it is

**CONTACT** Maoxin Zhang maoxin.zhang@cemo.uio.no Centre for Educational Measurement, University of Oslo, P.O. Box 1161 Forskningsparken, Oslo 0318, Norway

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10627197.2023.2222585

personal, meaning that individuals may handle the same problem differently; and (d) it consists of multiple processes (Mayer & Wittrock, 2006). Primarily four cognitive processes are relevant in problem-solving: *representing*, *planning/monitoring*, *executing*, and *self-regulating* (Mayer & Wittrock, 2006). *Representing* refers to converting a given external problem environment into an internal mental representation. As an example, respondents may need to explore the problem environment by conducting activity-based manipulations to better understand and represent the problem (Jonassen, 2000). *Planning* occurs when respondents devise a means of achieving the goal, whereas *monitoring* refers to an evaluation of the effectiveness and appropriateness of the planned solution (Mayer & Wittrock, 2006). Given that planning is a resource-intensive mental activity, researchers have chiefly employed information from response times to infer planning (e.g., Albert & Steinberg, 2011; Eichmann, Goldhammer, Greiff, Pucite, & Naumann, 2019). *Executing* means carrying out the solution as planned while *self-regulating* refers to investigation and modification of the solution, such as checking previous actions, taking remedial actions, and potentially starting the problem over (Schunk, 2003).

## Process data analysis

Many assessments involving problem-solving have been implemented on computers, producing a new type of data named process data. Process data include the whole human-computer interactive process and give a record of everything that a test-taker did through the course of the assessment. Depending on the assessment environment, this can for example be keystrokes, mouse clicks, and the timestamps for each such operation. These highly detailed data provide researchers with valuable information that can improve the understanding of test-takers' cognitive processes (OECD, 2014a) and analyses of these data benefit educational measurement. For instance, by analyzing process data, we can identify test-taking disengagement (Sahin & Colvin, 2020), profile students (Gnaldi, Bacci, Kunze, & Greiff, 2020; Greiff, Molnár, Martin, Zimmermann, & Csapó, 2018), improve the measurement precision of problem-solving proficiency (Han, Liu, & Ji, 2022; Liu, Liu, & Li, 2018), and validate the interpretation of test scores (Ercikan & Pellegrino, 2017).

However, it is challenging to analyze and make use of process data due to their special characteristics. First, different test-takers experience distinct cognitive processes and differently interact with the computer, leading to varied lengths of process data. By comparison, a standardized test normally consists of a fixed number of items and responses. Second, process data contain a great amount of noise (Tang, Wang, He, Liu, & Ying, 2020). Furthermore, the dependencies of sequential operations in process data pose challenges to directly applying conventional psychometric models, such as item response theory, factor analysis, and regression models.

To discover the underlying processes based on event logs, researchers have primarily adopted two approaches. The first approach focuses on the automatic construction of models that aim to reproduce all the observed operations via algorithms such as the fuzzy miner (Bogarín, Cerezo, & Romero, 2018). In another example, Hanga, Kovalchuk, and Gaber (2020) applied recurrent neural networks to model the relationships among all events and predict the next event by using a process map to visualize the probability of the event transitions. Such methods depict the response processes well, but establishing connections to problem solving theory is not straightforward. Hidden Markov models (HMMs) can also describe the complete sequence of operations and identify latent stochastic states underlying the observed operation sequences. HMMs have been applied to process data to understand response strategies by examining the latent states and the transition between the latent states (Xiao, He, Veldkamp, & Liu, 2021). However, the latent states are highly task-dependent and the interpretation of different latent states is often difficult to ascertain.

Different from directly analyzing the complete event sequence, the second approach is to reduce the dimension of process data by extracting statistics (called features) and subsequently conduct analyses with the extracted features. Several feature extraction methods have been considered in the literature. Yuan, Xiao, and Liu (2019), for example, defined a set of behavioral indices according to the

Assessment and Teaching of 21$^{st}$ Century Skills project (Griffin & Care, 2014) from a theory-driven perspective. Such indices are easy to interpret but require substantial effort in re-coding procedures and are not easily applied to different types of problems. By comparison, data-driven methods automatically extract features via techniques from machine learning and natural language processing (e.g., Qiao & Jiao, 2018; Tang, Wang, Liu, & Ying, 2020). A widely-used method taken from natural language processing is the analysis of n-grams where the complete operation sequence is decomposed into smaller units, such as single operations (unigrams) and operation sequences with two (bigrams) or three (trigrams) consecutive operations (He & von Davier, 2016). Different tasks would lead to different grams, making the conclusions hard to generalize and compare across multiple tasks. In addition, dissimilarity measures have also been applied to capture the discrepancy of the operation sequences between individuals via multidimensional scaling methods (Tang, Wang, He, Liu, & Ying, 2020) and between pre-defined optimal sequences via the Longest Common Subsequence algorithm (Hao, Shu, & von Davier, 2015; He, Borgonovi, & Paccagnella, 2021; Ulitzsch et al., 2021). Similarity and efficiency indicators are then defined to indicate the extent to which respondents follow an optimal strategy and the extent to which respondents conduct redundant actions (He, Borgonovi, & Paccagnella, 2021). However, such approaches require defining optimal strategies, have not considered that respondents may combine multiple strategies (He, Borgonovi, & Paccagnella, 2021), and can be computationally demanding (Ulitzsch et al., 2021). Furthermore, researchers have borrowed techniques from fields such as network analysis (Newman, 2010) to represent and analyze process data. A more detailed introduction to network analysis is presented in the next subsection.

Although response times and operation sequences can be simultaneously logged with computer-based problem-solving tasks, most studies have centered on operation sequences and have had less focus on response times. However, some studies have included time-on-task (e.g., Greiff, Niepel, Scherer, & Martin, 2016; Vörös & Rouet, 2016) and, recently, researchers have incorporated response times in operation sequence analysis (e.g., Chen, Li, Liu, & Ying, 2019; Ulitzsch et al., 2021; Xu, Fang, & Ying, 2020). Since response times can reflect test-takers' mental processes, we aim to also utilize detailed information from response times via an approach of network analysis in this study.

### Network analysis approach

In this subsection, we introduce the approach of network analysis. Networks are combinations of points (called "vertices") connected by directed or undirected lines (called "edges") (Newman, 2010). Networks are widely used in varied fields. For instance, a social network can represent the relationships among a group of people, where the vertices represent people and the edges the friendships between people. Another example is the World Wide Web with vertices representing web pages and edges representing hyperlinks. That is, the meanings of vertices and edges depend on the application.

The rationale to apply network analysis to process data is that operations are directly connected with the previous and the next operation. Researchers have represented process data via a network with vertices representing operations and edges representing the transition of the operations (Vista, Awwal, & Care, 2016; Zhu, Shu, & von Davier, 2016). Specifically, Vista, Awwal, and Care (2016) treated a set of operation sequences as a directed network and proposed an exploratory network analysis to investigate the topology of dominant parts of the network. The dominant parts were determined based on the importance of the vertices and edges. In another study, Zhu, Shu, and von Davier (2016) weighted the edges with the frequency and computed the statistics *weighted density* (indicates how dense the network is), *degree centrality* (indicates how important the vertex is), and *reciprocity* (indicates the mutual relationship between two vertices). Moreover, they investigated many triadic patterns (the relationships among three vertices) in their process data. These studies showcased the potential of applying network analysis to process data. Different from previous studies that focused on the local patterns of dominant paths (Vista, Awwal, & Care, 2016) or triadic relationships (Zhu, Shu, & von Davier, 2016), this article emphasizes the general features of the complete network based on different topic areas related to problem-solving theory.

### Present study

In this article, we introduce an approach for identifying problem-solving solution patterns via techniques of network analysis and Gaussian mixture models. Our study has three aims. First, we describe a method for defining networks from response times and operation sequences. Second, we define and extract relevant network features from each participant's process data which reflect the valuable and aggregated information of problem-solving processes. Third, with the extracted features, we classify participants into distinct solution pattern groups with Gaussian mixture models and present their typical solution patterns. The remainder of the article is organized as follows. The methods section describes how we visualize process data with networks, how we define network features, and introduces Gaussian mixture models. Thereafter, a case study is presented to illustrate the application of the proposed method with a real data set. We finally conclude with a discussion.

## Methods

### Network visualization of process data

We employ weighted directed networks to visualize process data in this study. Namely, we denote operations as vertices and the sequence of two successive operations as a directed edge. The edges are then weighted by the time spent on the transition. This distinguishes our study from previous studies that used the frequency of transitions as weights (Zhu, Shu, & von Davier, 2016). For a general representation of a network, we denote a set of vertices in the network as $V = \{v_1, v_2, \cdots, v_N\}$, where $N$ denotes the number of all the possible operations; a set of edges in networks as $E = \{e_1, e_2, \cdots, e_m\}$, where $m$ denotes the number of edges in the network, which is potentially different for each test-taker; and a set of weights on the edges as $T = \{t_1, t_2, \cdots, t_m\}$, which corresponds to the response times spent on the transitions.

To illustrate the approach, consider the following process data example. In the example, there are six possible type-specific interactions (A, B, C, D, E, F) in addition to two system-defined interactions (Start and End). Assume that a test-taker took a set of operations (operation sequence: {Start, A, B, C, A, D, F, D, End}) and that the corresponding response times in seconds were {5, 3, 2, 1, 2, 3, 4, 2}. The operations and response times were restructured into an edge list in Table 1 and into a network graph in Figure 1. In this edge list, each row represents a certain transition from one operation to another. The first two columns (i.e., the "From" and "To" columns) refer to the operations and the third column indicates the time spent on each transition. If a certain edge occurred more than once, the average time spent on this transition was used as the weight of the edge. From Figure 1, we can get a more straightforward visualization of how the operations are linked. First, for the vertices, there is an isolated vertex labeled E in the graph, meaning that this interaction was not conducted by this test-taker. In contrast, some of the vertices were extensively connected with other vertices, such as vertex

**Table 1.** An example of edge list.

| From | To | RT |
|---|---|---|
| Start | A | 5 |
| A | B | 3 |
| B | C | 2 |
| C | A | 1 |
| A | D | 2 |
| D | F | 3 |
| F | D | 4 |
| D | End | 2 |

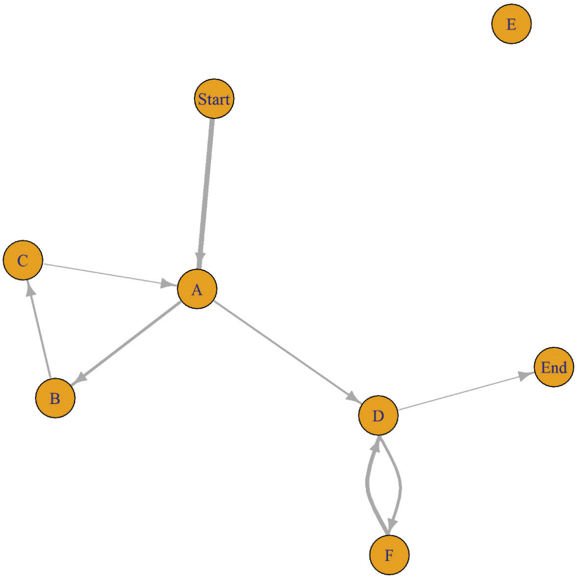


**Figure 1.** A network representation of process data, where the strength of the edges is weighted by the average response times.

A and D. Second, for the edges, we can have an overview of the transitions of operations. In addition, it is easy to visualize certain transitions of operations. For example, the connections between vertex D and F were mutual, which indicated that the test-taker conducted operation D and revisited it after operation F. Such edges are called *reciprocated edges*. Additionally, the relationships among three vertices are widely discussed in network analysis. For example, in Figure 1, vertices A, B, and C have a transitive relationship because they are interconnected with each other. Thus, such a combination of three vertices is called a *transitive triad*. Meanwhile, the relationship among vertices Start, A, and D is called a *structural hole* where one vertex is connected with two other vertices but where those two vertices are not connected. Third, since the edges were weighted by the response times, a thicker edge indicates that more time was spent on the transition. The variations among the thicknesses of edges indicate how the test-taker distributed the time on the varied operations. In this example, the test-taker spent more time at the starting stage (indicated by the thickest edge from Start to A) and less time on other stages. For each respondent, the network graph can clearly illustrate the individual process data. However, it becomes challenging to distinguish and compare many network graphs using visual inspection. Instead, we need to extract informative statistics from the networks. The next subsection introduces the network features that we define from the network of process data.

### Network features of process data

In deciding which network features to use, we set out to define variables that reflect the cognitive processes of problem solving. First, considering the *representing* process, respondents can explore the task environment by conducting specific operations (OECD, 2014a). To reflect the exploring aspect of response behavior, we defined vertex features since vertices correspond to operations. Second, since time-related measures have been proposed to infer the cognitive process of *planning/monitoring* (e.g., Eichmann, Goldhammer, Greiff, Pucite, & Naumann, 2019), we weight the transitions of operations by the response time. In a previous study by Eichmann, Goldhammer, Greiff, Pucite, and Naumann (2019), the longest duration and the variance of the durations have been proposed to measure planning in dynamic problems where not all the necessary information is presented at the outset (Stadler, Niepel, & Greiff, 2019). In the current study, we extract similar time-related measures from the network of process data. Third, the *executing* process depends on specific actions being taken in sequence (Mayer & Wittrock, 2006) and, hence, the sequence of operations is of high interest with respect to this process. Therefore, edge features are introduced as a way to capture information about sequential operations. To better describe the relationships between vertices, we break down the relationships into multiple levels: between two operations, among three operations, between two categories of operations (i.e., correct/incorrect operations), and among the complete network. Last, with respect to the *self-regulating* process, we consider if respondents check and modify the solution by revisiting previous actions. The edge features regarding two or three operations thus reflect the *self-regulating* process. In sum, we proposed network features (vertex features, edge features, and time-related features) that strongly relate to the cognitive processes in problem solving. Specifically, we define seven network features (see Table 2).

## Vertex features

*Operation diversity* is defined as the proportion of the number of present operations to the number of all possible operations. A higher value of operation diversity indicates that the test-taker had more diverse interactions with the computer and is indicative of exploration behavior (OECD, 2014a).

## Edge features

The edge features describe the relationships among the operations, reflecting the *executing* and *self-regulating* processes (Mayer & Wittrock, 2006). We organize the description of edge features according to which vertices they concern. To be specific, we introduce the relationships among all vertices (*edge density*), between two vertices (*reciprocity*), among three vertices (*transitivity*), and among specific groups of vertices (the *External – Internal index*). We define *edge density*, the fraction of the possible edges that are actually present in the network (Hanneman & Riddle, 2005), to indicate the extent to which a respondent performed transitions among the operations. We also consider *reciprocity* and *transitivity*, which focus on the dyadic and triadic relationships in the network, respectively. The *reciprocity* of a network is the proportion of the number of reciprocated edges to the number of all

**Table 2.** The definitions and interpretations of the seven network features.

| Feature | Formula | Notes | Interpretations |
|---|---|---|---|
| *Operation diversity* | $n/N$ | $n$ and $N$ are the number of non-isolated and total number of vertices, respectively. | Indicates if the test-taker had diverse interactions with the computer. |
| *Edge density* | $m/(n * (n-1))$ | $m$ is the number of edges existing in the network. | Captures the extent to which the test-taker tended to perform transitions among the existing operations. |
| *Reciprocity* | $k/m$ | $k$ is the number of reciprocated edges. | A high value indicates that the test-taker tended to revisit previous operations. |
| *Transitivity* | $TT/(TT+SH)$ | $TT$ and $SH$ refer to the number of transitive triads and structural holes. | A high value indicates that the test-taker tended to revisit previous operations after conducting one additional operation. |
| *External-Internal index* | $(Co - Inco)/(Co + Inco)$ | $Co$ and $Inco$ denote the number of correct edges (pointing to correct operations) and incorrect edges, respectively. | Measures whether the test-taker took correct operations constantly. |
| *Average time* | $\bar{t} = \sum_{i=1}^{m} t_i/m$ | $t_i$ is the response time spent on the $i^{th}$ edge. | Measures the average time spent on the transitions. |
| *Standard deviation of time* | $\sqrt{\sum_{i=1}^{m} (t_i - \bar{t})^2/(m-1)}$ | $\bar{t}$ is the average time. | Reflects whether the test-taker evenly distributed the time spent on the transitions. |

existing edges in the network. The *transitivity* feature is defined as the ratio of the number of transitive triads over the total number of transitive triads and structural holes. Both *reciprocity* and *transitivity* capture if the respondent tended to revisit previous operations, which are relevant to the cognitive process of *self-regulating*. If a network has a high level of *reciprocity* and *transitivity*, it indicates that the individual conducted operations back and forth; on the contrary, if the values are zero, it indicates that the individual conducted operations straight ahead and never revisited previous operations. Additionally, we employ the *External – Internal (E-I) index* (Krackhardt & Stern, 1988) to measure whether test-takers took correct operations constantly. Note that correct operations are defined by researchers on an item-by-item basis. The calculation of the *E-I index* is the difference between the number of correct edges and incorrect edges over the number of total edges, ranging from −1 to 1. A higher level of the *E-I* index implies that the test-taker conducted more correct operations than incorrect operations. Hence, the *E-I index* to some extent refers to the correspondence between a respondent's solution and the optimal solution, and is thus similar to the efficiency indicator defined in He, Borgonovi, and Paccagnella (2021).

### *Time-related features*

The last two features focus on the distribution of response times; namely, *average time* and *standard deviation of time* (*sd of time*). The former indicates how much time was spent on the transitions of operations on average. The latter shows the variation of response times; namely, whether the test-taker distributed the time equally on the transitions or not. The time-related features can be viewed as measures of the cognitive process of *planning/monitoring* (Eichmann, Goldhammer, Greiff, Pucite, & Naumann, 2019).

## *Gaussian mixture models*

After extracting features from individual networks, we aim to discover hidden subgroups on the basis of network features and summarize the common response pattern within each subgroup. Test-takers in the same subgroup should be similar to each other but distinct from test-takers in other subgroups (James, Witten, Hastie, & Tibshirani, 2013). The question arising immediately is whether there exist substantially different subgroups or clusters. We assess the clustering tendency to answer the question via the *Hopkins statistic* (Lawson & Jurs, 1990), which examines the spatial randomness of the data. To be specific, a sample with elements $x_i$ is randomly drawn from a real dataset $X$; next, a simulated sample with elements $y_i$ is generated from a uniformly distributed dataset that has the same variation as the dataset $X$; then the distance of both $x_i$ and $y_i$ with its nearest neighbor in $X$ is computed and denoted as $w_i$ and $u_i$ respectively. The *Hopkins statistic* is then defined as $\quad u_i/(\Sigma u_\mathrm{i} + \quad w_i)$. If the dataset $X$ includes meaningful clusters, the distance for the real data points $x_i$ should be much smaller than the distance for the artificial data points $y_i$ leading to a higher value for the *Hopkins statistic*.

Next, to obtain the group membership for each test-taker, we apply Gaussian mixture models (GMMs; Fraley & Raftery, 2002). Mixture models belong to latent variable modeling and the latent variable is assumed to be discrete instead of the continuous latent variable in factor analysis. In GMMs, a multivariate Gaussian distribution of the observed variables for each cluster is assumed. Namely, different clusters are centered at a distinct mean vector and the covariance matrix across clusters can have different geometric features such as volume, shape, and orientation. To estimate the GMMs we used the expectation-maximization (EM) algorithm with model-based hierarchical agglomeration as initialization (Fraley & Raftery, 2002). Subsequently, two crucial questions arise – the number of clusters and the geometric features of the covariance matrix. In model-based clustering, we run a number of possible models and select a final model according to certain criteria. Information criteria such as Bayesian Information Criterion (BIC; Schwarz, 1978) and the integrated complete-data likelihood criterion (ICL; Biernacki, Celeux, & Govaert, 2000) have been widely used, which take both model fit and model complexity into consideration. Another approach is the bootstrap likelihood ratio test (LRT), which compares a mixture model with *k* clusters and another mixture model with *k + 1* clusters via a resampling approach to obtain LRT significance (McLachlan, 1987). In short, we consider multiple possible models with a cluster size ranging from one to nine and where all possible covariance structures are considered. Our final model was selected based on information-based and resampling-based criteria. A complete procedure will be illustrated in the following section.

## Network analysis of problem-solving tasks in PISA 2012

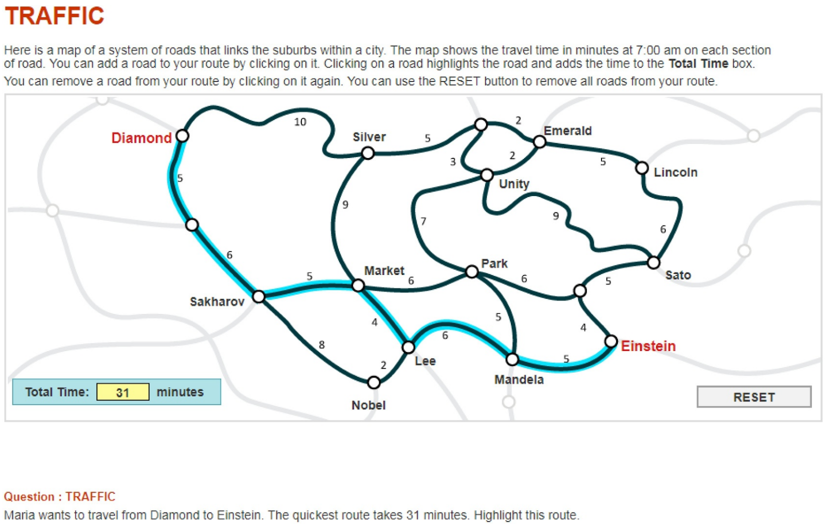
### *Data and sample*

#### *Task*

In this article, we illustrate the application of the proposed approach with a creative problem-solving task (the traffic task, see Figure 2) in PISA 2012, which aimed to assess students' problem-solving competence – the capacity to engage in, understand, and resolve problems when the solution is opaque (OECD, 2014a). The traffic task presents a map with 23 paths and the travel time of each path to the test-taker. The object is to find the quickest route from Diamond to Einstein and the shortest possible time is provided. Test-takers can activate or deactivate each path by clicking on it and they can also reset the map by clicking on the "RESET" button.

#### *Participants*

We used data from the United States. The data were retrieved from http://www.oecd.org/pisa/data. These are anonymized secondary data and neither consent to participate or consent for publication nor ethics approval were required for the current study. There were 413 students participating in the traffic task, but seven of them only conducted one operation and they were excluded from this analysis. Afterwards, there were 406 participants in our analysis: 191 of them were females and 76% of the participants solved the problem successfully.



TRAFFIC

Here is a map of a system of roads that links the suburbs within a city. The map shows the travel time in minutes at 7:00 am on each section of road. You can add a road to your route by clicking on it. Clicking on a road highlights the road and adds the time to the **Total Time** box.

You can remove a road from your route by clicking on it again. You can use the RESET button to remove all roads from your route.

Question : TRAFFIC

Maria wants to travel from Diamond to Einstein. The quickest route takes 31 minutes. Highlight this route.

Figure 2. Traffic task in the domain of creative problem-solving in PISA 2012, where the highlighted route is the solution. The figure is retrieved from https://www.oecd.org/pisa/test –2012/testquestions/question2/.

### Log-file data and re-coding

The PISA 2012 creative problem-solving tasks were delivered by computer. We used the sequence of operations and the corresponding timestamps from the log files (OECD, 2014b). The sequence of operations mainly consists of two parts: system-defined operations (start and end item) and task-specific operations (hit_path, reset). An example of a task-specific operation is "hit_DiamondSilver," which means clicking on the path between Diamond and Silver, as illustrated in Figure 2. Since there are substantial differences between selecting and de-selecting a path, we distinguished "hit_path" operations as "select_path" and "cancel_path." In addition, there were operations beyond the instructions of the task, such as clicking on the timer box or the paragraph above the map. Most studies ignored these operations because they contributed nothing to solving the problem. However, we noticed that such operations were quite common and believe these operations also provide us with useful information. Hence, we denoted all these operations as a new category of operation: an insignificant operation. In this way, the operations include 50 types: start, end, reset, insignificant operation, 23 path selections, and 23 path cancellations. To simplify the name of the paths, we denoted them from P1 to P23 (see Appendix A for details). P1 - P6 are required in the correct solution, while the remaining paths should be excluded. Regarding the timestamps, we computed the time difference between two successive operations as the response time. So far, we have reorganized the data and can move on to the network analysis of process data.

### Example networks and descriptive statistics

The reduced process data were restructured for each student into an edge list as indicated in Table 1. Then, we plotted the network graph of process data for each student via the "igraph" package (Csardi & Nepusz, 2006) in R 4.0.2 (R Core Team, 2020). The example R code used in this study can been found in Appendix B. An example network is presented in Figure 3. In this graph, we colored the vertices according to their relationships with the correct solution: golden = correct selection (SP1 - SP6), red = correct cancellation (CP7 - CP23), blue = incorrect cancellation (CP1 - CP6), incorrect selection (SP7 - SP23) and insignificant operations, and gray = other operations (start, end and reset). In addition, the size of each vertex is determined by the number of edges connected with it. In other words, a larger vertex means that the operation was conducted more frequently. From Figure 3, we can see that this test-taker actively interacted with the computer. The individual conducted both correct and incorrect operations in the process and finally found the correct solution. This participant preferred canceling the paths one by one to using the reset button. Sometimes, the test-taker conducted insignificant operations. According to the thickness of the edges, this individual spent more time on the first and the last transition.

### Gaussian mixture models based on network features

The next procedure is to identify solution patterns of the respondents. Given that test-takers who solved the problem successfully or unsuccessfully may exhibit distinct patterns, we separated the test-takers into a success group and a failure group according to their final performance on the task and then applied GMMs to them separately. We normalized *Sd of time* by the logarithmic function and then assessed the clustering tendency of the data. The *Hopkins statistic* was .899 and .906 in the failure and success groups, respectively, which were larger than the cutoff value .75 (Lawson & Jurs, 1990) and indicated that meaningful patterns existed in the data. Next, we used the "mclust" package (Scrucca, Fop, Murphy, & Raftery, 2016) to estimate GMMs in the failure and success groups. The final model was selected based on either of the BIC, ICL, or bootstrap LRT. All selection procedures suggested the same models: a two-cluster mixture with covariances having the same volume but distinct shapes and orientations (EVV) in the failure group and a four-cluster mixture with varying volumes, shapes, and orientations (VVV) in the success group.

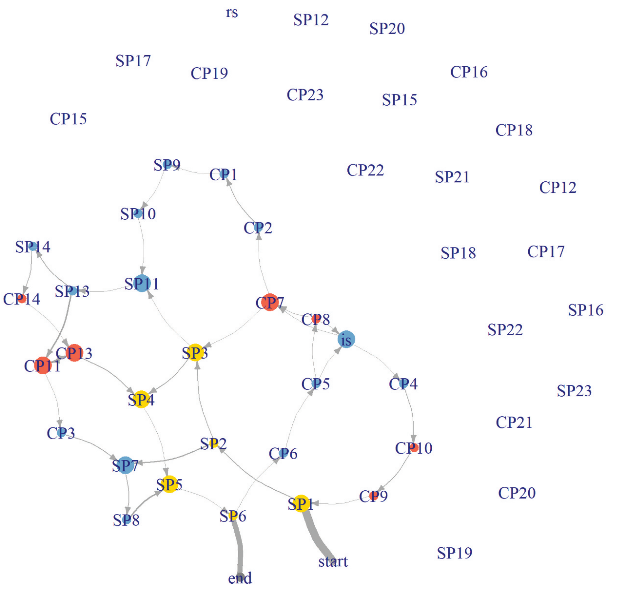


**Figure 3.** A network representation of process data (studentID = 02778) in the traffic task. rs = reset. is = insignificant operations. Color illustration for vertices: golden = correct selection; red = correct cancellation; blue = incorrect operations and the insignificant operation; gray = start, end, and reset.

Subsequently, we present the results from the selected GMMs. In Table 3, we present descriptive statistics of the seven network features. Additionally, we summarize the plausible values from the problem-solving, mathematics, and reading domains for each cluster because student problem-solving performance is related to the performance of mathematics and reading (OECD, 2014a), which potentially provide evidence for the validity of the clustering results. Plausible values describe the performance of the population (OECD, 2014b), with a mean equal to 500 and standard deviation equal to 100 across all countries. There were five sets of plausible values for each domain. Following Mislevy, Beaton, Kaplan, and Sheehan (1992), we aggregated the plausible values and estimated the mean and its standard error for each cluster in Table 3. We also plot the cluster-level graphs in Figure 4 by aggregating operation sequences of students in the same cluster. To make the cluster-level graphs clearer, we ignored the transitions that occurred less than 10% of the cluster size.

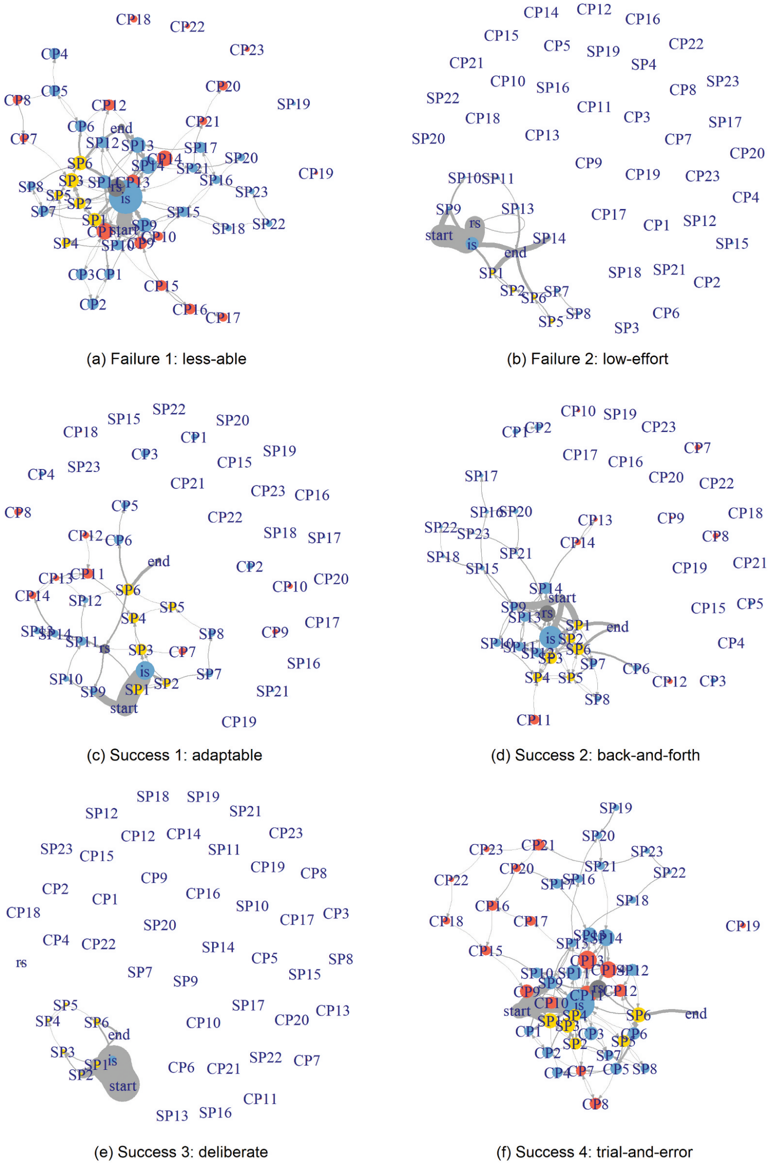
Combining the information from the cluster-level network graphs and network features, we then seek to interpret the common solution pattern in each cluster. We first focus on the failure group. Failure 1 actively interacted with the computer and tried various operations (*operation diversity* = .571) with relatively short (*average time* = 2.330) but even time (*sd of time* = 1.027) spent on the transitions. Students in this cluster tended to revisit previous operations; however, they did not conduct correct operations constantly (*E-I index* = .117). We denoted Failure 1 as the less-able cluster. They engaged with the task and made a great effort but might lack knowledge or the ability to learn from errors. Members in Failure 1 also performed below average in the three proficiency domains (see Table 3). In contrast, although the average time on transitions was quite long (*average time* = 10.541) in Failure 2, it seems that these students just stayed in the system but did not engage with the problem (*operation diversity* = .142). According to Figure 4b, students in Failure 2 tended to try one incorrect solution (i.e., {SP1, SP2, SP7, SP8, SP5, SP6} or {SP9, SP10, SP11, SP13, SP14}) and then gave up. We denoted this cluster as the low-effort cluster. Students in this cluster had the lowest value of plausible values, implying that these students might display a lack of effort in the whole cognitive assessment.

We subsequently describe the success groups. From Figure 4c, students in Success 1 first tried an incorrect solution as students in Failure 2 did. However, instead of giving up quickly, students in Success 1 attempted to modify their solutions by de-selecting the incorrect paths (e.g., CP7 and CP11), minimizing the difference between their solution and the optimal solution. Without too many trials, they were able to find the correct solution. They spent longer and more uneven time on transitions than Success 2 and Success 4, indicating that they planned more for the operation sequences that they executed. Consequently, their *E-I index* was higher than that of Success 2 and Success 4 (see Table 3). We denoted Success 1 as the adaptable cluster in which students performed well in the three proficiency domains. Students in Success 2 conducted more diverse operations than students in Success 1. More importantly, they were likely to conduct operations back-and-forth, indicated by a high level of *reciprocity* and *transitivity* in Table 3. Namely, they tended to revisit previous operations, and thus we named this cluster the back-and-forth cluster. The third cluster in the success group, Success 3, showed a highly distinct pattern. Specifically, students in this cluster did not conduct many operations (*operation diversity* = .182), but the majority of the operations were required in the correct solution according to Figure 4e, resulting in a high value for the *E-I index* (.956). Another striking characteristic of this cluster is that they spent a long time on the transitions and that the standard deviation of response times was huge (see Table 3). Figure 4e helps explain the results. There it can be seen that students spent quite a large proportion of time at the starting stage (indicated by rather thick edges pointing from "start"). After that, the students just spent a small amount of time on other operations. This indicates that the students first made a plan for the task, which requires a remarkably longer time, and then executed the plan straight ahead (indicated by low *reciprocity* and *transitivity*) and efficiently. Hence, we called this the deliberate cluster. Students in this cluster had a similar mean score in the problem-solving domain to the other clusters but the means in mathematics and reading competency were lower than students in other success clusters. The last cluster in the success group is quite similar to Failure 1 except for a higher level of the *E-I index* in Table 3.

**Table 3.** Means and standard errors of network features and plausible values in each cluster.

| Features | Failure 1 | Failure 2 | Success 1 | Success 2 | Success 3 | Success 4 |
|---|---|---|---|---|---|---|
| Size | 73 | 26 | 106 | 54 | 34 | 113 |
| *Operation diversity* | .571 | .142 | .369 | .420 | .182 | .668 |
| | (.025) | (.009) | (.009) | (.013) | (.007) | (.015) |
| *Edge density* | .064 | .195 | .068 | .081 | .117 | .050 |
| | (.003) | (.019) | (.002) | (.002) | (.002) | (.001) |
| *Reciprocity* | .076 | .132 | .004 | .118 | .022 | .058 |
| | (.008) | (.037) | (.001) | (.013) | (.008) | (.006) |
| *Transitivity* | .120 | .049 | .023 | .119 | .011 | .115 |
| | (.010) | (.024) | (.004) | (.013) | (.008) | (.006) |
| *External-Internal index* | .117 | −.001 | .533 | .362 | .956 | .335 |
| | (.039) | (.066) | (.021) | (.030) | (.021) | (.014) |
| *Average time* | 2.330 | 10.541 | 3.049 | 2.280 | 11.730 | 2.124 |
| | (.108) | (2.476) | (.134) | (.078) | (1.541) | (.072) |
| *SD of time* | 1.027 | 7.071 | 1.473 | .961 | 8.762 | .905 |
| | (.010) | (.366) | (.012) | (.005) | (.230) | (.006) |
| *PRO PV* | 467 | 420 | 542 | 521 | 523 | 542 |
| | (9.639) | (13.425) | (7.412) | (9.232) | (14.248) | (7.227) |
| *MAT PV* | 475 | 438 | 526 | 520 | 504 | 521 |
| | (9.107) | (12.642) | (7.589) | (9.386) | (12.311) | (7.899) |
| *REA PV* | 495 | 442 | 539 | 530 | 508 | 538 |
| | (9.253) | (14.197) | (6.842) | (9.755) | (14.049) | (6.771) |

PRO = problem-solving. MAT = mathematics. REA = reading. PV = plausible value.


CP18 CP22 CP23 CP4 CP5 CP20 CP8 CP12 SP19 CP21 CP7 CP6 end SP12 SP13 SP17 CP14 SP20 SP6 SP14 SP21 CP19 SP3 SP11 CP13 SP16 SP8 rs SP23 SP5 is SP2 SP7 SP15 SP1 SP18 SP22 SP9 CP11 start CP10 SP4 SP10 CP9 CP3 CP1 CP15 CP2 CP16 CP17
CP14 CP12 CP16 CP15 CP5 SP19 CP22 CP21 SP4 CP10 SP16 CP8 SP23 SP22 CP11 CP18 CP3 SP17 CP13 SP20 CP7 CP20 CP9 CP19 CP23 SP10 SP11 CP4 SP9 SP13 CP17 rs CP1 SP12 start SP15 is SP14 end SP18 SP21 SP1 CP2 SP2 SP7 SP6 SP8 CP6 SP5 SP3
SP22 SP15 SP20 CP18 CP1 CP3 SP19 CP15 SP23 CP4 CP21 CP23 CP16 CP5 CP8 CP22 SP18 SP17 CP12 CP6 end CP2 CP13 CP11 CP20 SP6 CP10 CP14 SP12 SP5 SP4 CP17 SP13 SP14 SP8 CP9 SP11 rs SP3 CP7 is SP16 SP7 SP10 SP2 SP9 SP1 SP21 start CP19
CP10 SP19 CP1 CP2 CP23 CP7 CP17 CP16 SP17 CP20 CP22 SP16 SP20 CP13 CP9 CP18 SP22 SP23 CP8 CP14 SP18 SP21 CP19 CP21 SP15 SP14 start CP15 CP5 SP9 SP13 rs SP1 end SP10 is SP2 SP11 SP12 SP6 CP4 SP3 SP7 CP6 SP4 SP5 CP12 CP3 SP8 CP11
SP18 SP19 SP12 SP21 CP12 CP14 CP23 SP23 CP15 SP11 CP9 CP19 CP8 CP2 CP16 SP10 CP1 CP18 CP17 CP3 SP20 CP4 CP22 SP14 CP5 SP8 rs SP7 SP9 SP15 SP5 SP17 CP13 SP4 SP6 CP10 CP20 end SP3 SP1 is SP22 CP7 SP2 CP6 CP21 start CP11 SP13 SP16
SP19 SP20 CP23 CP21 SP23 SP21 CP20 SP22 CP22 SP17 SP16 SP18 CP16 CP18 CP17 CP19 SP13 SP14 CP15 SP15 CP13 SP10 SP11 CP14 SP12 SP9 rs CP12 CP9 CP11 CP10 is start SP6 end SP1 SP4 SP3 CP1 CP3 SP5 CP6 SP2 CP2 SP7 CP4 CP7 CP5 SP8 CP8

(a) Failure 1: less-able

(b) Failure 2: low-effort

(c) Success 1: adaptable

(d) Success 2: back-and-forth

(e) Success 3: deliberate

(f) Success 4: trial-and-error

**Figure 4.** Cluster-level network representations of process data in the traffic task. rs = reset. Is = insignificant operations. Color illustration for vertices: golden = correct selection; red = correct cancellation; blue = incorrect operations and the insignificant operation; gray = start, end, and reset.

Success 4 tried various (*operation diversity* = .668) operations quickly (*average time* = 2.124) until the correct solution was found. Their solution pattern was in line with a trial-and-error strategy (Klahr, 2002). Hence, this cluster was denoted as the trial-and-error cluster. They were persistent to solve the problem, indicating high motivation in the assessment. In addition, their performances in the three domains were above average. So far, we have investigated the cluster-level networks of the problem-solving task and provided interpretations of the results, advancing the insight into students' solution patterns.

## Discussion

Process data from computer-based assessments provide researchers with valuable sources of inferring test-takers' mental processes. However, analyzing process data is often challenging. In this article, we represented operation sequences and response times from process data as network graphs, providing a straightforward way to visualize process data. In addition, we defined specific network features to extract useful information from the process data and shed light on the cognitive processes in problem solving. Based on the network features, we then identified solution patterns for success and failure groups through GMMs. A case study was conducted to showcase the approach.

### *Discussion on the empirical study*

With process data of the PISA 2012 traffic task from 406 students in the United States, we identified two and four clusters for the failure and success groups, respectively. We interpreted these clusters as the less-able, low-effort, adaptable, back-and-forth, deliberate, and trial-and-error clusters. Some reflections can be made based on our results. The behavior of students in the adaptable cluster is similar to the problem-solving technique means-ends analysis (Simon & Newell, 1971). This technique involves students first looking for differences between the present state and the desired state and then applying operators to reduce the differences. Such a strategy commonly occurs in human problem-solving behavior (Simon & Newell, 1971). Similarly, trial-and-error is also popular in problem-solving (Klahr, 2002), even though it is inefficient for complicated problems. Note that the less-able cluster seemingly adopted a similar approach but they were more likely to get stuck in incorrect operations and eventually failed to complete the task. By comparison, a more efficient solution pattern is manifested in the deliberate cluster where students first made a plan and then executed it straight ahead. It has been documented that prior planning plays an important role in problem-solving, but also that this effect varies across tasks (Eichmann, Goldhammer, Greiff, Pucite, & Naumann, 2019). For instance, in dynamic problems, test-takers cannot make a thorough plan at first because not all the information is provided at the outset. There were both static and dynamic tasks in PISA 2012. Generally applying a planning strategy at first does not guarantee good performance for dynamic tasks. We note that the deliberate cluster had only an average level of performance in the general problem-solving domain and other domains. This could indicate that the students who applied a planning strategy for the traffic task also, unsuccessfully, tried the same strategy for other tasks. Note that the traffic task under study represents a specific type of problem – the shortest path search problem, which has wide application to car navigation. To efficiently solve the problem, many algorithms have been proposed of which the bi-directional algorithm (Noto & Sato, 2000) is an example. It proceeds by finding the optimal path from both the starting point and the terminal point, which is similar to the strategy employed by students in the back-and-forth cluster. In short, the six clusters showed distinct solution patterns, and students in the clusters performed differently in the proficiency domains. By analyzing the response patterns we have gained insights into test-takers' problem-solving processes which go beyond what the binary outcome (success or failure) can possibly provide. However, the validation of the results and interpretations needs additional investigation.

Table 4. Conditions for applying network features.

| Feature | Free number of operations | Allowance of repeated operations | Operation categories | Free time allocation |
|---|---|---|---|---|
| *Operation diversity* | × | - | - | - |
| *Edge density* | × | × | - | - |
| *Reciprocity* | - | × | - | - |
| *Transitivity* | - | × | - | - |
| *External-Internal index* | - | - | × | - |
| *Average time* | - | - | - | × |
| *Standard deviation of time* | - | - | - | × |

Note. "×" indicates the property of tasks needed to employ particular network features. "-" indicates the property of tasks not needed to employ particular network features.

## Discussion on the proposed approach

Given that process data has become commonplace in many types of assessment, we believe there exists great potential to apply the proposed approach in diverse areas and fields. Once unique operations, operation sequences, and timestamps have been identified, the network based on the process data can be defined. However, a decision must be made regarding which network features to utilize and this depends on the research intentions and the type of task under analysis. In this article, we defined seven general network features and it is worthwhile to consider the suitable conditions for when to utilize the features. Here, we discuss the type of tasks and the conditions for applying the proposed features in Table 4. Four properties of tasks are relevant to the current study: (a) if the task requires a fixed number of operations or respondents are free to conduct any operations; (b) if the task allows respondents to conduct the same operations multiple times; (c) if the operations can be categorized into different groups (e.g., correct/incorrect); and (d) if respondents can freely distribute their time on the operations. Based on these properties, we then discuss under which conditions we can apply the network features. *Operation diversity* applies to the situation in which test-takers are free to conduct any operation. We take the *traffic* task as a viable example where test-takers can decide which paths to select or de-select. However, in other tasks, test-takers could be asked to conduct a fixed number of operations, resulting in the same *operation diversity*. *Edge density* is applicable when the number of operations is free and operations can be repeatedly conducted. Next, analysts can compute *reciprocity* and *transitivity* when test-takers are allowed to revisit previous operations. The *E-I index* can be used when analysts categorize operations into different types. Last, the time-related network features are meaningful when test-takers are allowed to allocate their time freely. Given respondents can usually conduct operations and distribute time freely, we believe that these network features have great potential in extracting valuable information from process data in future empirical studies.

Since there have emerged plenty of studies using process data, a brief comparison between the proposed approach and existing relevant methods with a focus on investigating response processes is described here. Compared to algorithms like fuzzy nets (Günther & van der Aalst, 2007) which reflect the results of process discovery models (e.g., the probability of taking a specific action subsequently), our approach reflects the original operation sequence and emphasizes features from networks. Additionally, educational process mining algorithms can automatically construct process discovery models, whereas the network features are pre-defined to reflect the cognitive processes of problem solving, making the results more interpretable. Compared to HMMs (e.g., Xiao, He, Veldkamp, & Liu, 2021), where solution patterns based on latent states are inferred, we directly identified different solution patterns from observed operations and response times. Compared to studies based on n-grams that capture short-length sequences (He & von Davier, 2016), it is easier to grasp complex relationships among the operations with network analysis by considering for example the density of the complete network. With the possible operations rising, the number of grams, especially bi-grams and tri-grams, will increase to a great extent, making the analysis and interpretations more complex. In contrast, an increasing number of possible operations would still correspond to only seven network features in our approach. Compared with the methods based on dissimilarity measures (e.g., He, Borgonovi, & Paccagnella, 2021) that particularly focus on the relationship between individual operation sequence and the optimal sequence, we considered the *E-I index* to reflect the extent to which respondents conducted necessary operations. Besides the *E-I index*, we also consider other aspects of operations and response times that are not closely related to the optimal strategies. Last, researchers have also utilized process data to infer respondents' problem-solving ability (e.g., Han, Liu, & Ji, 2022; Shu, Bergner, Zhu, Hao, & von Davier, 2017; Zhan & Qiao, 2022). However, in the current study, we have focused on respondents' solution patterns and estimation of problem-solving ability is beyond the scope of this article. In addition, we did not draw a conclusion about stronger and weaker problem-solvers among respondents who solve the task successfully.

### Contributions and limitations

This article has several contributions. First, we incorporate two types of data sources from the log-files of computer-based assessments – operation sequences and response times – in a single study, while the majority of the studies on process data analysis (e.g., He, Borgonovi, & Paccagnella, 2021; Zhu, Shu, & von Davier, 2016) ignored the response times. Second, networks of process data provide direct visualization of the whole problem-solving process. Compared to the raw process data, the seven network features that we defined can reduce the complexity but retain the essential information in the data. Additionally, the features are easy to obtain due to the availability of user-friendly tools such as the igraph package (Csardi & Nepusz, 2006) in R. The network features utilized in this study are based on the global networks instead of the local networks (e.g., Vista, Awwal, & Care, 2016), making them easier to generalize to multiple tasks.

With respect to the empirical study, we argue that identifying solution patterns is beneficial for students, teachers, and test development. Students can review their own network of process data, which helps them reflect on their cognitive processes, and teachers can tailor their instructions for each student after identifying the solution patterns of students. For instance, for students that exhibit a low-effort behavior pattern, teachers can try to identify why they did not make an effort in the assessment. Two such reasons could be a lack of basic computer skills to interact with the computer or that a low-stake test did not motivate them. For less-able students, it would be beneficial to give them easier exercises and offer explicit hints. For the deliberate group, planning everything at first might not be the most useful strategy in dynamic problems. Rather, exploration behavior can play a key role in such contexts (Eichmann, Greiff, Naumann, Brandhuber, & Goldhammer, 2020) and teachers can encourage students with the deliberate solution pattern to explore more within the system. Regarding the back-and-forth pattern, besides the bi-directional algorithm in the shortest path search problem (Noto & Sato, 2000) and the *self-regulating* process, researchers also found a similar pattern in other types of tasks (e.g., Xiao, He, Veldkamp, & Liu, 2021) and interpreted it as hesitation or uncertainty, which provides further information for teachers. For the trial-and-error and adaptable clusters, they were engaged with the test and performed quite well and as such would not require explicit instruction from teachers. Concerning test development, comparing test-takers' solution patterns with the intended task design is useful for the validation of the interpretations of task performance. In summary, the proposed approach incorporates operation sequences and response times and delineates how test-takers solve problems, which has practical significance in education.

Some limitations of our study should be noted. First, the seven network features cannot reflect all the information available in process data. We only considered certain types of global network features in this study. However, there are potentially other network features that can be utilized in the network of process data, such as the *centrality* of each operation to indicate the importance of individual operations, if single operations are of particular interest. In addition, using average response times as edge weights does not entirely capture the frequency of single action transitions (Zhu, Shu, & von Davier, 2016) and the variation of response times for single action transitions, although we weighted the vertices (e.g., operations) with associated frequencies in graphs. Deciding on what network features to compute depends on the specific study at hand and future studies can explore additional features beyond those considered here. Second, although we have shed light on the cognitive processes involved in problem-solving, it is not straightforward to create one-to-one connections between cognitive processes and network features. The cognitive processes occur in parallel (Lesh & Zawojewski, 2007) and the network features can reflect several cognitive processes. Third, validating the cluster interpretations remains a challenge. In our study, we have provided supporting evidence for the approach used via the computation of the average domain scores in each cluster. However, there are no external response patterns that can directly confirm our cluster results due to the limitations of the present analysis. Qualitative methods such as interviews and think-aloud protocols can offer some added evidence regarding the cluster interpretations for future studies with a different design and data collection method. Additionally, there exist graph mining techniques that aim to find the frequent sub-graphs and discover topological structures from a geometry-oriented perspective (Bogarín, Cerezo, & Romero, 2018; Ulitzsch et al., 2021), which could be a potential direction for future studies. Last, like other studies using single tasks, it is a question that whether the solutions patterns found in the study can be generalized to other tasks. The problem type (e.g., dynamic or static), task setting, and difficulty may influence respondents' solution patterns. For example, the back-and-forth pattern would not apply to tasks that forbid revisiting previous operations. If the generalization of the solution patterns is of interest, it is possible to employ the proposed approach to each task separately, compare the cluster-level network features, and examine the extent to which the cluster members of one task are also grouped to the same cluster in other tasks.

## Conclusions

In this article, we introduced an approach that combines network analysis and Gaussian mixture models to visualize process data as a network, extract network features from process data, and identify problem-solving solution patterns. A real data set from PISA 2012 was used to illustrate the complete procedure and we gained a deeper understanding of how students solve the problem beyond the binary final performance. Our results indicate that mining information embedded in process data provides an insight into the cognitive processes of students. In addition, our proposed approach demonstrates great potential in analyzing process data and exploring solution patterns in problem-solving tasks in practice.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Maoxin Zhang http://orcid.org/0000-0002-6250-3610
Björn Andersson http://orcid.org/0000-0002-9007-2440

## References

Albert, D., & Steinberg, L. (2011). Age differences in strategic planning as indexed by the tower of London. *Child Development*, *82*(5), 1501–1517. doi:10.1111/j.1467-8624.2011.01613.x

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725. doi:10.1109/34.865189

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1), e1230. doi:10.1002/widm.1230

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, *10*, 486. doi:10.3389/fpsyg.2019.00486

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *Inter Journal, Complex Systems*, *1695*(5), 1–9.

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, *128*, 1–12. doi:10.1016/j.compedu.2018.08.004

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning*, *36*(6), 933–956. doi:10.1111/jcal.12451

Ercikan, K., & Pellegrino, J. W. (2017). Validation of score meaning using examinee response processes for the next generation of assessments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 1–8). New York, NY: Routledge.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631. doi:10.1198/016214502760047131

Gnaldi, M., Bacci, S., Kunze, T., & Greiff, S. (2020). Students' complex problem solving profiles. *Psychometrika*, *85*(2), 469–501. doi:10.1007/s11336-020-09709-2

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, *126*, 248–263. doi:10.1016/j.compedu.2018.07.013

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46. doi:10.1016/j.chb.2016.02.095

Griffin, P., & Care, E. (2014). Assessment and teaching of 21st century skills: Methods and approach. Dordrecht, the Netherlands: Springer. doi:10.1007/978-94-017-9395-7

Günther, C. W., & van der Aalst, W. M. (2007). Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In *International conference on business process management* (pp. 328–343). Springer: Berlin, Heidelberg. September.

Hanga, K. M., Kovalchuk, Y., & Gaber, M. M. (2020). A graph-based approach to interpreting recurrent neural networks in process mining. *IEEE Access*, *8*, 172923–172938. doi:10.1109/ACCESS.2020.3025999

Han, Y., Liu, H., & Ji, F. (2022). A sequential response model for analyzing process data on technology-based problem-solving tasks. *Multivariate Behavioral Research*, *57*(6), 960–977. doi:10.1080/00273171.2021.1932403

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. Riverside, CA: University of California Riverside .

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining*, *7*(1), 33–50.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, *166*, 104170. doi:10.1016/j.compedu.2021.104170

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749–776). Hershey, PA: Information Science Reference. doi:10.4018/978-1-4666-9441-5.ch029

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer New York. doi:10.1007/978-1-4614-7138-7

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research & Development*, *48*(4), 63–85. doi:10.1007/BF02300500

Klahr, D. (2002). *Exploring science: The cognition and development of discovery processes*. Cambridge, Massachusetts: The MIT Press.

Krackhardt, D., & Stern, R. N. (1988). Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*, *51*(2), 123–140. doi:10.2307/2786835

Lawson, R. G., & Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences*, *30*(1), 36–41. doi:10.1021/ci00065a010

Lesh, R., & Zawojewski, J. (2007). Problem solving and modeling. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 763–802). Reston, Estados Unidos: National Council Teachers of Mathematics.

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology*, *9*, 1372. doi:10.3389/fpsyg.2018.01372

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. Alexander, P. Winne, & G. Phye (Eds.), *Handbook of educational psychology* (pp. 287–303). Mahwah, NJ: Erlbaum.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *36*(3), 318–324. doi:10.2307/2347790

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161. doi:10.1111/j.1745-3984.1992.tb00371.x

Newman, M. (2010). *Networks: An introduction*. New York: Oxford University Press.

Noto, M., & Sato, H. (2000). A method for the shortest path search by extended Dijkstra algorithm. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Nashville, TN, USA, *3*, 2316–2320.

OECD. (2014a). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V)*. Paris, France: OECD.

OECD. (2014b). *PISA 2012 technical report*. Paris, France: OECD.

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*, *9*, 2231. doi:10.3389/fpsyg.2018.02231

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education*, *8*(1), 1–24. doi:10.1186/s40536-020-00082-1

Schunk, D. (2003). Self-regulation and learning. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of psychology* (Vol. 7, pp. 59–78). New York: Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289–317. doi:10.32614/RJ-2016-021

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling*, *59*(1), 109–131.

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26* (2), 145–159. doi:10.1037/h0030806

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. *Intelligence*, *72*, 1–12. doi:10.1016/j.intell.2018.11.003

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, *85*(2), 378–397. doi:10.1007/s11336-020-09708-3

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2020). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, *74*(1), 1–33. doi:10.1111/bmsp.12203

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika*, *86*(1), 190–214. doi:10.1007/s11336-020-09743-0

Vista, A., Awwal, N., & Care, E. (2016). Sequential actions as markers of behavioural and cognitive processes: Extracting empirical pathways from data streams of complex tasks. *Computers & Education*, *92*, 15–36. doi:10.1016/j.compedu.2015.10.009

Vörös, Z., & Rouet, J.-F. (2016). Laypersons' digital problem solving: Relationships between strategy and performance in a large-scale international survey. *Computers in Human Behavior*, *64*, 108–116. doi:10.1016/j.chb.2016.06.018

Xiao, Y., He, Q., Veldkamp, B., & Liu, H. (2021). Exploring latent states of problem-solving competence using hidden Markov model on process data. *Journal of Computer Assisted Learning*, *37*(5), 1232–1247. doi:10.1111/jcal.12559

Xu, H., Fang, G., & Ying, Z. (2020). A latent topic model with Markov transition for process data. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 474–505. doi:10.1111/bmsp.12197

Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of collaborative problem solving based on process stream data: A new paradigm for extracting indicators and modeling dyad data. *Frontiers in Psychology*, *10*(369), 1–14. doi:10.3389/fpsyg.2019.00369

Zhan, P., & Qiao, X. (2022). Diagnostic classification analysis of problem-solving competence using process data: An item expansion method. *Psychometrika*, *87*(4), 1529–1547. doi:10.1007/s11336-022-09855-9

Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement*, *53*(2), 190–211. doi:10.1111/jedm.12107

Paper II

# Investigating planning and non-targeted exploration in PIAAC 2012: Validating their measures based on process data and investigating their relationships with problem-solving competency

II

**Maoxin Zhang, Björn Andersson, Samuel Greiff**

Published in *Journal of Intelligence*.

Journal of
*Intelligence*

MDPI

*Article*

# Investigating Planning and Non-Targeted Exploration in PIAAC 2012: Validating Their Measures Based on Process Data and Investigating Their Relationships with Problem-Solving Competency

**Maoxin Zhang [1,*], Björn Andersson [1] and Samuel Greiff [2]**

[1] Centre for Educational Measurement, Faculty of Educational Sciences, University of Oslo, 0316 Oslo, Norway; bjorn.andersson@cemo.uio.no
[2] Department of Behavioral and Cognitive Sciences, University of Luxembourg, L-4366 Esch-sur-Alzette, Luxembourg; samuel.greiff@gmail.com
* Correspondence: maoxin.zhang@cemo.uio.no

**Abstract:** Problem-solving is a critical aspect of intelligence that has become increasingly important in modern society. Mapping out the determinants of success in problem-solving helps understand the underlying cognitive processes involved. This article focuses on two key cognitive processes in problem-solving: non-targeted exploration and planning. We generalize previously defined indicators of planning and non-targeted exploration across tasks in the 2012 Programme for the International Assessment of Adult Competencies and examine the internal construct validity of the indicators using confirmatory factor analysis. We also investigate the relationships between problem-solving competency, planning, and non-targeted exploration, along with the specific dependence between indicators from the same task. The results suggest that (a) the planning indicator across tasks provides evidence of internal construct validity; (b) the non-targeted exploration indicator provides weaker evidence of internal construct validity; (c) overall, non-targeted exploration is strongly related to problem-solving competency, whereas planning and problem-solving competencies are weakly negatively related; and (d) such relationships vary substantially across tasks, emphasizing the importance of accounting for the dependency of measures from the same task. Our findings deepen our understanding of problem-solving processes and can support the use of digital tools in educational practice and validate task design by comparing the task-specific relationships with the desired design.

**Keywords:** log-file data; large-scale assessment; PIAAC; problem-solving; planning; non-targeted exploration

check for updates

**Citation:** Zhang, Maoxin, Björn Andersson, and Samuel Greiff. 2023. Investigating Planning and Non-Targeted Exploration in PIAAC 2012: Validating Their Measures Based on Process Data and Investigating Their Relationships with Problem-Solving Competency. *Journal of Intelligence* 11: 156. https://doi.org/10.3390/jintelligence11080156

Received: 3 April 2023
Revised: 30 June 2023
Accepted: 2 August 2023
Published: 7 August 2023

## 1. Introduction

In modern societies, solving problems is a major task in our life (OECD 2014), involving multiple higher-order cognitive skills such as devising plans, testing hypotheses, remedying mistakes, and self-monitoring (Greiff et al. 2015). Thus, a high level of problem-solving competency lays a sound foundation for future learning and prepares students to handle novel challenges (Csapó and Funke 2017; OECD 2014). To make students better problem-solvers, it has been suggested to explicitly embed problem-solving skills into national curricula (Greiff et al. 2014) and use computer-based problem-solving simulations called "microworlds" where students can explore and discover underlying rules and regulations (Ridgway and McCusker 2003). Besides acquiring problem-solving competency in formal education, it is also important to develop such a skill over the entire lifetime and engage in lifelong learning (Greiff et al. 2013). For example, teachers might need to learn how to employ digital tools for long-distance education, and office workers might need to adapt to a different computer system. It has been documented that proficiency in applying information and communication technology (ICT) skills to solve problems has a positive influence on participation in the labor force (Chung and Elliott 2015). That is, the competency of problem-solving is both a key objective of educational programs (OECD 2014) and valued in the workplace.

Hence, many educational large-scale assessments for students and adults have focused on the domain of problem-solving. For example, the Programme for the International Student Assessment (PISA) evaluated 15-year-old students' problem-solving in 2003, 2012, and 2015. Another example is the 2012 Programme for the International Assessment of Adult Competencies (PIAAC), which covers problem-solving in technology-rich environments when using ICT. Many of these assessments have been implemented on computers where the complete human–computer interactions are recorded in log files. Just as the task performance provides information on what respondents can achieve, the log files open a window into how respondents approach the task. Log files offer valuable information for researchers to understand respondents' cognitive processes when solving problems, and this study intends to explore the log files of problem-solving tasks to infer the cognitive processes when solving problems.

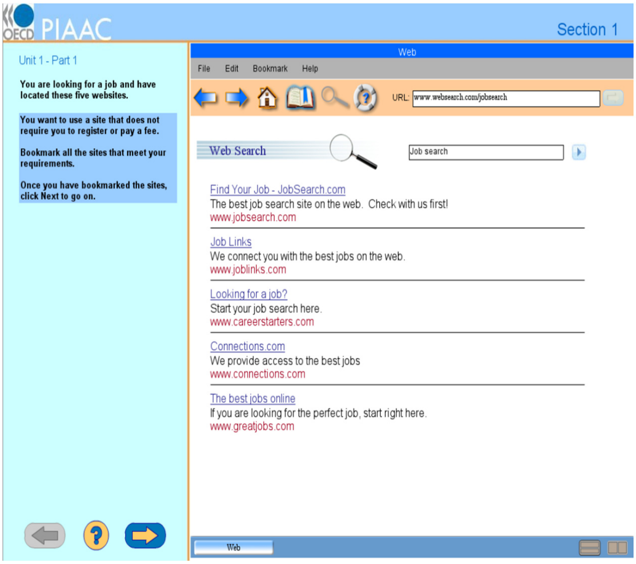
A better understanding of the problem-solving processes has potential implications for integrated assessments and learning experiences (Greiff et al. 2014). For example, the analysis results from log files can provide teachers with materials on the weaknesses and strengths of students in solving a problem, and further, teachers can tailor their instruction for students. In this study, we aim to improve the understanding of the cognitive problem-solving processes in the context of information processing. This can potentially benefit educational practices related to improving problem-solving skills. For example, the analysis of log files can inform teachers whether a student is engaged in solving a problem or applies an efficient strategy to approach the problem (Greiff et al. 2014) and whether additional instructional scaffolding is needed when a student is stuck.

The data availability of international large-scale assessments has stimulated studies that explore the information from the log files. Both theory-based methods (e.g., Yuan et al. 2019) and data-driven methods based on machine learning or natural language processing (e.g., He and von Davier 2016) have been applied to extract information called process indicators from log files, and the relationships between these process indicators and task performance have then been inferred. However, the majority of research has focused on single tasks, and the generalizability of the conclusions remains unclear. In this study, we used process indicators to analyze multiple tasks involving two cognitive aspects of problem-solving: planning and non-targeted exploration. Specifically, we examine the internal construct validity of the measures of planning and non-targeted exploration using tasks from PIAAC 2012 and infer their relationships with problem-solving competency. Next, we review the literature on problem-solving, planning, and non-targeted exploration and describe the current study in more detail.

### 1.1. Problem-Solving

A problem is considered to have two attributes: (a) the difference between a given state and the desired goal state and (b) the social, cultural, or intellectual worth embedded in achieving the goal (Jonassen 2000). Problems can be categorized into different types according to their characteristics. Here, we introduce three problem categories based on dynamics, structuredness, and domain (Jonassen 2000). First, problems can be categorized as static or dynamic problems based on the dynamics of a problem situation. In static problems, all the information relevant to the problem is known at the outset (Berbeglia et al. 2007). In contrast, dynamic problems (also called complex problems) do not present all the necessary information at the outset; instead, problem-solvers must interact with the problem situation to collect relevant information (Stadler et al. 2019). Thus, exploring the problem situation plays a more important role in dynamic problems compared with static problems. In addition, according to the structuredness (i.e., the clarity of a problem), a problem can be mapped into a curriculum with two poles representing well-structured and ill-structured problems (Arlin 1989). Problems in textbooks tend to be well-structured problems with a clearly defined initial and goal state and operator rules, whereas problems such as designing a building are ill-structured problems. The tasks in PISA 2012 and PIAAC 2012 are relatively well-structured problems, and the optimal solutions are predefined. Moreover, based on the specific domain knowledge required to solve a problem, problems can be categorized as domain-specific and domain-general (Jonassen 2000). For example, physics and biology exams typically present domain-specific problems. In contrast, finding a quickest route between two places and figuring out why a lamp is not working are examples of domain-general problems in everyday contexts.

The cognitive process of transferring a given state into a goal state when the solution is not immediately accessible is called problem-solving (Mayer and Wittrock 2006). Mayer and Wittrock (2006) argued that problem-solving involves several component processes: representing, planning/monitoring, executing, and self-regulating. We take a problem-solving task released from the PIAAC 2012 (see Figure 1) as an illustrative example. The task requires participants to bookmark job-seeking websites that do not need registration or fees. When confronted with this problem, respondents must convert the given information into a mental representation, which includes the initial state (e.g., five website links in this example), goal state (e.g., bookmarked websites satisfying the requirements), and the possible intermediate states (Bruning et al. 2004). Such a process is called representing. Planning occurs when respondents devise a way to solve the problem (Mayer and Wittrock 2006), such as decomposing it by checking the links from the first to the last to see which require registration or a fee. Monitoring refers to the process of evaluating whether the solution is valid and effective (Mayer and Wittrock 2006). Implementing the planned operations is called executing (Mayer and Wittrock 2006). Self-regulating involves modifying and maintaining activities that allow respondents to move toward the goal (Schunk 2003). While these processes are all assumed to be active in problem-solving, the importance of each cognitive process differs across problems.



Unit 1 - Part 1

You are looking for a job and have located these five websites.

You want to use a site that does not require you to register or pay a fee.

Bookmark all the sites that meet your requirements.

Once you have bookmarked the sites, click Next to go on.

www.websearch.com/jobsearch

Job search

Find Your Job - JobSearch.com
The best job search site on the web. Check with us first!
www.jobsearch.com

Job Links
We connect you with the best jobs on the web.
www.joblinks.com

Looking for a job?
Start your job search here.
www.careerstarters.com

Connections.com
We provide access to the best jobs
www.connections.com

The best jobs online
If you are looking for the perfect job, start right here.
www.greatjobs.com

**Figure 1.** An example task released from the PIAAC PS-TRE domain. The figure was retrieved from https://piaac-logdata.tba-hosting.de/public/problemsolving/JobSearchPart1/pages/jsp1-home.html (accessed on 19 October 2021).

In a technology-rich society, problems often appear because new technology is introduced (OECD 2012). On the other hand, tools and technologies are widely applied to facilitate problem-solving. Capturing the intersection of problem-solving competency and the skills needed in ICT, the 2012 PIAAC specifically covers a domain called problem-solving in technology-rich environments (PS-TRE), where problem-solving competency is defined as the capacity of "using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (OECD 2012, p. 47). The 2012 PIAAC PS-TRE domain developed fourteen problems that are dynamic, relatively well-structured, and domain-general information problems. The problems are assumed to assess a single dimension—problem-solving competency (OECD 2012). In addition to problem-solving competency, PIAAC 2012 also emphasizes the cognitive dimensions of problem-solving. The PS-TRE domain shares similar cognitive problem-solving processes with Mayer and Wittrock (2006) but with a particular focus on acquiring and dealing with information in computer-based artifacts.

To acquire the relevant information, it is necessary to interact with the problem environment and explore the features or potential resources that are closely related to the representing process. After collecting useful information, respondents may devise a plan (e.g., to break down the problem and set sub-goals for achieving the desired state). These two processes, exploration and planning, play vital roles in problem-solving and are thus the focus of this study. We next introduce the definitions and measures of planning and exploration (particularly non-targeted exploration) and their relationships with task performance.

### 1.2. Planning and Problem-Solving

Planning is defined as mental simulations of future operations and associated outcomes with the aim of achieving certain goals or guiding problem-solving (Mumford et al. 2001). An early conception of planning referred to certain predefined, fixed sequences of operations. More recently, however, researchers have argued that adaptable cognitive responses are at the core of planning (Mumford et al. 2001). In addition, it is assumed that planning consists of multiple and distinguishable processes (Hayes-Roth and Hayes-Roth 1979). For example, Mumford et al. (2001) proposed a planning process model: prior to developing an initial and general plan, environment analyses including the identification of resources and contingencies are necessary. Then, an initial plan needs to be elaborated into a more detailed plan, which requires searching information about potentially useful operations and resources needed to execute these operations (Xiao et al. 1997). Based on the forecasting of outcomes from these operations, one may refine the plan and then execute it.

Planning is a generative activity that is hard to observe directly. Early qualitative studies applied think-aloud protocols and content analyses to investigate planning (e.g., Xiao et al. 1997). Recently, quantitative measures have been used to facilitate research on planning, such as evidence from functional neuroimaging (Unterrainer and Owen 2006) and time-related measures (Albert and Steinberg 2011; Eichmann et al. 2019; Unterrainer et al. 2003). In this study, we consider the process measure of response times as an indicator of planning. Because planning is resource-intensive (Mumford et al. 2001), the time spent making a plan should be much longer than the time spent actually executing the plan. The time-related measures capture the quantity of planning. If a respondent rushes into a problem and randomly tries different operations until a correct solution is found (i.e., a trial-and-error strategy), the value of the time-related measures would be relatively small, indicating a small quantity of planning.

In the context of problem-solving, the time-related measures of planning differ between static problems and complex problems. A commonly used measure of planning in static problems, such as the Tower of London, is the first-move latency (Albert and Steinberg 2011; Unterrainer et al. 2003). This measure, also known as preplanning time, is defined as the time interval between the beginning of the problem and the first action a respondent takes. However, in complex problems, respondents need to explore the simulated environment to generate information before they are able to make a plan that takes into account all relevant aspects of the problem situation at hand. In line with this thinking, Eichmann et al. (2019) expanded the measure of planning in complex problems from the first-move latency to the longest duration between moves. Namely, the authors argued that planning can appear at any time during the course of complex problem-solving. They also acknowledged that the longest duration cannot cover the entire planning process but that the main planning activity is captured by this indicator. Research on planning in complex problems is quite limited, and Eichmann et al.'s (2019) work seems to be the first on this topic, thus, yielding important implications for the current study.

Planning is of interest not only because it is a cognitive process in problem-solving but also because it influences task success or task performance (Albert and Steinberg 2011; Eichmann et al. 2019). Theoretically, planning provides a mental model of the problem by identifying critical issues and relevant strategies and promotes optimized and effective solutions by organizing the chunks of operations (Mumford et al. 2001). However, previous empirical research showed diverse results regarding the relationship between task success and planning due to different types of problems and different indicators of planning. For instance, and as mentioned above, Albert and Steinberg (2011) found a positive relationship between first-move latency and task success in static problems, whereas Eichmann et al. (2019) did not find such an effect for the longest duration indicator in dynamic problems. Additionally, Eichmann et al. (2019) derived two other indicators of planning to describe the time taken before the longest duration appears (the delay indicator) and the variability in time intervals between two successive operations (the variance indicator). They found that planning in the early stages benefited task performance (i.e., a negative relationship between the delay indicator and task scores) and that a longer duration indicator in a later stage or continued planning activities could compensate for a lack of early planning. Their models implicitly indicate that each indicator from different tasks implies similar meanings (Assumption I) and that the relationships between the planning indicators and task success are consistent across tasks (Assumption II). However, we argue that these assumptions (i.e., Assumptions I and II) require explicit examination. In addition, although the random effects in their models captured the variances at the task level, the specific relationships between the indicators and task performance at the task level remained unaccounted for.

### *1.3. Non-Targeted Exploration and Problem-Solving*

To better understand the nature of the problem, test-takers need to explore the problem environment (e.g., navigate through different computer interfaces or pages) to uncover new information. Exploration refers to behaviors that investigate and seek information that is beyond the instructions of the task (Dormann and Frese 1994). Some exploratory behaviors are goal-oriented (goal-directed behaviors), leading to achieving a desired goal state. On the other hand, some exploratory behaviors can be irrelevant to solving the problem (non-targeted behaviors), such as clicking on some buttons on the interface to check their functions and exploring some pages that do not contain useful information for the problem (Eichmann et al. 2020a, 2020b). Note that both goal-directed and non-targeted behaviors help test-takers understand the problem but in different ways. Goal-directed behaviors capture the relevant points and convey similar information as task success because the problem cannot be successfully solved without these goal-directed behaviors, whereas non-targeted behaviors provide additional information compared to task success.

One research field related to non-targeted exploration is error management, where errors are defined as unintended deviations from goals (Frese et al. 1991). It is found that compared to participants who received step-by-step guidance on programming (i.e., error avoidance or goal-directed exploration), participants who were encouraged to explore the system, make mistakes, and learn from them (i.e., non-targeted exploration) during the training stage performed better during the testing stage (Frese and Keith 2015). One explanation is that non-targeted exploration plays a role in representing the problem (Eichmann et al. 2020b; Kapur 2008). Test-takers who were encouraged to explore the environment, in spite of making more errors, gained a better understanding of the problem setting, the potential features, and resources of the interfaces. In addition, participants who received more training on exploratory error management showed a higher level of metacognitive activity such as hypothesis-testing and monitoring (Keith and Frese 2005).

In computer-based problems, exploration is operationalized as human–computer interactions that refer to all the operations that respondents conduct in the computer system and are recorded in log files, such as mouse clicks and keyboard input. For each item, test developers and content experts have predefined one or more optimal solutions consisting of a minimum number of operations that can successfully solve the problem and thus represent the most efficient strategies (He et al. 2021). We can broadly categorize individual operations into goal-directed or non-targeted operations, depending on whether the operation is required to solve the problem or not (Eichmann et al. 2020a, 2020b). Goal-directed operations refer to operations that must be performed to solve the problem, which are operationalized as the operations that occur in any of the optimal solutions. In contrast, non-targeted operations are operations that are unnecessary to solve the problem, which are operationalized as the operations that do not occur in any optimal solutions. For example, in the task of Figure 1, clicking on and bookmarking the websites that satisfy the task requirements are goal-directed operations. However, clicking on the Help button in the menu is non-targeted because it is not included in the optimal solution.

Although non-targeted operations do not directly contribute to successful task completion (i.e., not occurring in any optimal solutions) and can appear erroneous, they have been found to benefit task performance (Dormann and Frese 1994), learning (Frese and Keith 2015), transfer performance (Bell and Kozlowski 2008), and meta-cognition (Bell and Kozlowski 2008). Eichmann et al. (2020a) also found that the number of non-targeted explorations is positively related to problem-solving competency, and the effects are consistent across 42 countries using the PISA 2012 problem-solving domain. The authors argued that non-targeted explorations facilitate goal-directed behaviors. Consider the Help button as an example. Although the Help button is not considered as a necessary operation to solve the problem, it provides test-takers with information about the functions of the menu, such as the function of the bookmark button, which can help test-takers better understand the potential resources in the computer system. When test-takers find the websites that meet the task requirements, they would know how to bookmark the websites.

A further aspect of defining an operation is whether it is performed for the first time or repeated. Implementing an operation for the first time is associated with information generation, whereas performing the same operation again indicates information integration (Wüstenberg et al. 2012). As a result, Eichmann et al. (2020b) distinguished between initial and repeated operations. Once a respondent performed a specific operation, such as clicking on the Help button in the task in Figure 1, the individual was assumed to gain information related to the Help button. If the respondent performed the same operation again, there would be little new information added to the problem space. Since exploration greatly concerns generating new information (Dormann and Frese 1994), we propose the number of initial non-targeted operations as a measure of the latent variable: non-targeted exploration. This differentiates our study from Eichmann et al. (2020b), who focused on both initial and repeated non-targeted operations.

### *1.4. The Current Study*

Previous studies by Eichmann and coauthors have deepened the understanding of planning and non-targeted exploration based on the PISA 2012 tasks (Eichmann et al. 2019, 2020a). However, the extent to which we can apply their definitions of planning and non-targeted exploration to the PIAAC 2012 information problems and the extent to which the indicators measure the same constructs require further research. If there is insufficient evidence of internal construct validity, it would be problematic to apply this measure to different items or different samples. Therefore, validating the internal construct of planning and non-targeted exploration across items is a crucial component of this study. We concurrently utilize information from multiple tasks and validate the approach of Eichmann and coauthors by looking at a more diverse set of tasks (i.e., PS-TRE) with a different population, namely, adults.

Furthermore, most studies analyzing process data of problem-solving tasks have only used log data from a single item (e.g., Ulitzsch et al. 2021; Chen et al. 2019), meaning the generalizability of the findings to other tasks is lacking. For example, it is an open question whether or not respondents apply similar strategies (e.g., trial-and-error) across tasks. Similarly, are the relationships between planning and problem-solving competency stable across tasks or are the relationships task-dependent? If the relationships are generalizable, then researchers and practitioners can use the findings across similar tasks. In this study, we examine the general and task-specific relationships between planning, non-targeted exploration, and problem-solving competency.

Our first set of research questions concerns the internal construct validity of the indicators for planning, non-targeted exploration, and problem-solving competency. If we find evidence that the same operationalization (see detailed definitions in Section 2.3) of the indicators is applicable across different items within different contextual settings, this implies that the indicators measure the same construct, thus providing support for internal construct validity for the indicators. Specific to the current study, we examine the construct validity of planning (*Q1a*), non-targeted exploration (*Q1b*), and problem-solving competency (*Q1c*) using a set of tasks from the PIAAC 2012 PS-TRE domain. For each item, we extract the indicators for planning, non-targeted exploration, and problem-solving competency along the same rationale. To examine evidence of construct validity, we applied confirmatory factor analysis (CFA; Jöreskog 1969) to each type of indicator. In CFA models, multivariate data are analyzed with the hypothesis that a latent variable underlies the observed variables (Bartholomew et al. 2011, p. 2). For example, the item response score is considered to be the observed indicator of the latent variable problem-solving competency. If the variations of the indicators across items can be adequately attributed to a latent variable, we can claim that the internal construct validity is established (AERA 2014).

The second set of questions that we are interested in points to the problem-solving competency's relationship with planning (*Q2a*) and non-targeted exploration (*Q2b*). Although previous studies have investigated such questions (e.g., Albert and Steinberg 2011; Unterrainer et al. 2003), only limited studies have examined the findings in dynamic problems (Eichmann et al. 2019, 2020b). Given that dynamic problems are becoming more popular in educational assessments and that the planning and exploration processes might differ between static and dynamic problems, examining their relationships with problem-solving competency is relevant and needed. In the research of Eichmann et al. (2019), the overall relationship between planning and task performance across tasks was examined, whereas if such a relationship might differ between tasks was uncounted for. Tasks differ in complexity, the interface, and the amount of information (OECD 2013), implying that the importance of planning and non-targeted exploration varies among the tasks. Hence, besides the overall relationships between the latent variables (i.e., planning, non-targeted exploration, and problem-solving competency), we also consider their task-specific relationships by adding residual correlations of observed indicators for planning, non-targeted exploration, and problem-solving competency from the same task. The variance of the errors can be attributed to individual differences among participants, task characteristics, and measurement error. The residual correlations that we added account for the additional dependence between indicators based on the same task, beyond the dependence induced by the correlations between the main factors of planning, non-targeted exploration, and problem-solving competency. Hence, by answering *Q2a* and *Q2b* from the levels of both latent variables and observed variables, we can gain a more fine-grained understanding of the research questions than Eichmann et al. (2019, 2020a). For *Q2a*, we hypothesized that the overall relationship between planning and problem-solving competency is negligible but that the relationship at the observed variable levels can be task-dependent, based on the results from Eichmann et al. (2019) and the diversity of tasks. For *Q2b*, because non-targeted exploration helps represent the problem and acquire information from available resources, we hypothesized a positive relationship between problem-solving competency and non-targeted exploration. Similarly, task-dependent relationships are also expected for *Q2b* because tasks differ in the extent to which respondents are allowed to interact with the interfaces. To achieve answers for *Q2a* and *Q2b*, we included all three indicators in a single model and considered the dependencies among the latent variables (i.e., the overall relationships) and the pairwise residual correlations of the three indicators from the same task (i.e., task-dependent relationships).

## 2. Materials and Methods

### *2.1. Participants and Tasks*

This study uses the performance data and associated log files from the 2012 PIAAC assessment. PIAAC is a program for assessing and analyzing adult skills and competencies that are essential to personal and societal success (OECD 2013). The stimuli materials were developed based on everyday life activities, and the target population was noninstitutionalized residents between 16 and 65 years of age in the country regardless of citizenship or language (OECD 2013). The PIAAC assessment was implemented by 25 countries (OECD 2012). All participating countries produced their sample design under the guidance of the PIAAC Technical Standards and Guidelines. In general, probability-based sampling methods were adopted to select an unbiased, randomized, and representative sample of the target population (OECD 2013). Countries developed their own sampling frames according to national situations. For example, Singapore had a full list of residents in the population registry that was used as a qualified sampling frame, and the sample was randomly selected based on the population registry. However, many countries like the United States adopted a multi-stage sampling method since such population registries did not exist there. In short, geographic domains such as provinces or states and dwelling units were randomly selected in primary stages, and persons in the domains had an equal probability to be sampled at the last stage of selection. After obtaining a sample, checks were conducted to ensure that the sample met the sampling plan. For example, the noncoverage rate of the target population was computed to indicate the portion of the target population not covered by the sample frames. In the United States, people who live in large, gated communities are not covered, and the noncoverage rate is 0.1%, which is the lowest in all participating countries (OECD 2013). For a more detailed description of the sampling design, readers are directed to the PIAAC technical report (OECD 2013). To avoid cultural heterogeneity and render the analyses of the vast log-file data manageable, we used only data from the United States. We chose the sample from the United States because of the low noncoverage rate, high response rates, and the large proportion of participants in the PS-TRE domain.

The 2012 PIAAC PS-TRE domain covers dynamic information problems that include one or more digital scenarios (e.g., email, web, word processor, and spreadsheet). Each PS-TRE task includes two panels (see Figure 1): The left panel shows the instructions that describe the scenario and the goal state (i.e., bookmarked websites fulfilling some requirements), and the right one represents the initial problem environment that corresponds to the given state. Respondents may need to first explore the system by, for example, clicking on the menu or a link to get to know the problem environment and then spend a relatively long time devising a plan to solve the problem. There are two booklets in PS-TRE, and each consists of seven fixed-order tasks. Based on the assessment design, test-takers randomly received zero, one booklet, or two booklets. We used the second booklet (PS-TRE2). Only participants with sufficient ICT skills in the background questionnaire had access to the PS-TRE tasks. Sufficient ICT skills include knowing how to manipulate the mouse and keyboard, understanding concepts like files and folders, and having experience with basic computer operations like save, open, and close files (OECD 2013).

### *2.2. Data Preparation*

The log files of the 2012 PIAAC domains can be downloaded from the GESIS Data Catalogue (OECD 2017). There were 1355 American participants in PS-TRE2, but 30 of them directly skipped all seven tasks and were excluded from the current analysis. The raw log files were preprocessed via the PIAAC LogDataAnalyzer (LDA) tool. The reformatted log data consisted of the following variables: respondent ID, item information, event_name, event_type (e.g., START, TOOLBAR, TEXTLINK), timestamp in milliseconds, and event_description, which describes the specific event (e.g., "id=toolbar_back_btn" means clicking on the back button in the toolbar). We recoded the data by filtering the system logs and aggregating the keyboard input and clicks in pop-up windows. A detailed explanation of this procedure is provided in Appendix A.

### *2.3. Measures*

For each student on each item, we extracted three indicators: task scores, longest duration, and the number of initial non-targeted operations, from performance data and the log files. In this subsection, we describe the three measures in detail.

*Problem-solving competency.* The indicators for problem-solving competency were response scores that can be extracted from the OECD website. In PS-TRE2, three items were scored dichotomously, and four were scored polychotomously by PIAAC. If a participant spent less than five seconds on a task, the response was scored as missing (OECD 2012). In the current data set, only five response scores were denoted as missing values by PIAAC. We directly used their scoring as the measures for the construct problem-solving competency.

*Planning.* We used the time intervals between consecutive events from log files to compute the longest duration, excluding the time interval for the last two events. The last two events are always NEXT_INQUIRY (request the next task) and END (end the task) based on the task design, and the intervals for the last two operations indicate reflection on the executed actions rather than planning. A simulated operation sequence and associated time intervals for the job-seeking task are presented in Table 1. Excluding the time intervals for the last two operations, we identified the longest one—10 s—as the longest duration indicator. For those who directly skipped a task, the longest duration was coded as missing. In a previous study, Eichmann et al. (2019) specified three indicators of planning: the longest duration, the variance indicator, and the delay indicator. However, we found the Pearson correlations between the indicators were around 0.80 for the PS-TRE tasks, and the longest duration typically occurred just after the task began, which meant that the delay indicator was often identical to the duration indicator. That is, the three aspects of planning from Eichmann et al. (2019) largely overlapped in our data, and we therefore used only a single planning indicator per item for the construct planning in this study.

**Table 1.** A simulated example of operation sequence and response times.

| Operation | Notes | Time Interval | Planning Indicator | Exploration Indicator |
|---|---|---|---|---|
| START | Enter the problem system | - | - | System-defined |
| textlink_page1 | Click on the first link | 10 s | Yes | IniNT |
| toolbar_back_btn | Click on the back button in the toolbar | 3 s | No | IniNT |
| web_menu_help | Click on the Help button in the menu | 5 s | No | IniNT |
| textlink_page5 | Click on the fifth link | 8 s | No | GD |
| toolbar_bookmark_btn | Click on the bookmark button in the toolbar | 7 s | No | GD |
| bookmark_add_page5 | Confirm adding the fifth page to bookmark | 4 s | No | GD |
| web_menu_help | Click on the Help button in the menu | 3 s | No | RepNT |
| NEXT_INQUIRY | Request the next task | 12 s | - | System-defined |
| END | End the task | 4 s | - | System-defined |

Note: IniNT = initial non-targeted. RepNT = repeated non-targeted. GD = goal-directed. We shortened the names of the operations in the raw log files.

*Non-targeted exploration*. To define the non-targeted exploration indicators, we first identified the unique operations for each task based on the log files of the participants. There were on average 200 unique operations (range = [57, 446]) in each of the PS-TRE2 tasks. Operations that occurred in any of the optimal solutions were considered goal-directed operations and the others non-targeted operations. Thereafter, we defined the indicator of non-targeted exploration as the number of initial non-targeted operations for each item. For the Figure 1 example, we supposed that the correct solution was {START, textlink_page5, toolbar_bookmark_btn, bookmark_add_page5, NEXT_INQUIRY, END}. By subsequently checking whether a given operation in Table 1 was included in the optimal solution, we identified goal-directed or non-targeted operations. The number of initial non-targeted operations, which was three in this example, served as the indicator of non-targeted exploration. For those who directly skipped a task, the indicator was coded as missing.

*Data transformation.* Latent variable modeling like factor analysis for continuous data (Jöreskog 1969) normally has the assumption of multivariate normality, but both process indicators (i.e., longest duration and the number of initial non-targeted operations) deviated from normal distributions according to large skewness and kurtosis (see Appendix B), requiring data transformation. One approach is the Box–Cox transformation (Box and Cox 1964). However, such one-to-one transformations do not work well when the data have many identical values (Peterson and Cavanaugh 2019). In addition, there are some extreme outliers in the longest duration and the number of initial non-targeted operations. Instead of transforming the indicators into normally distributed variables, we used quantiles to recode the process indicators into equal-sized categorical variables, which can reduce the impact of the outliers. Specifically, if the raw value was zero, we kept the value as it was; for the remaining values, we recoded the values as 1, 2, 3, and 4 with the 25%, 50%, and 75% quantiles as the cutoff values. Higher categories indicate that more initial non-targeted operations were applied, or a respondent spent more time planning than other respondents. In the following analysis, we treat the three types of indicators (response scores, longest duration, and the number of initial non-targeted operations) as ordered categorical data.

### 2.4. Analysis Procedures

In this study, we apply latent variable models to analyze the process indicators and task performance. Latent variable models are widely used in social sciences when researchers intend to measure a conceptual construct (Bartholomew et al. 2011) such as problem-solving competency. However, since it is difficult to measure the construct directly, researchers instead develop instruments based on theory to infer the construct indirectly. In PIAAC 2012, a battery of items was developed to measure problem-solving competency, and respondents' responses to the test are collected and considered as observed indicators of the unobserved construct (i.e., problem-solving competency). In analyzing the observed responses, the researchers extract what is common in the indicators. The latent variable that explains the common variability of the observed indicators is then interpreted as the problem-solving competency afterward. A similar approach is used to measure the latent variables of planning and non-targeted exploration, where the longest duration and the number of initial non-targeted operations from multiple items are used as observed indicators, respectively.

To answer the research questions related to the internal construct validity (i.e., *Q1a*/*Q1b*/*Q1c*), we applied confirmatory factor analysis (CFA; Jöreskog 1969) to each type of indicator. CFA is widely used to examine the latent construct by specifying the relationships between the observed indicators and latent variables on the basis of specific hypotheses (Brown 2015). We hypothesized that latent planning would underlie the longest duration (Model 1a), latent non-targeted exploration would underlie the number of initial non-targeted operations (Model 1b), and latent problem-solving competency would underlie the observed task scores (Model 1c). That is, the latent variables govern the associated observed indicators and thus explain the common variability of the indicators. To test these hypotheses, we examine if the hypothetical models fit well with the real data by checking the goodness-of-fit of the models and factor loadings that inform on the relationship between the observed indicators and the latent variable.

Regarding *Q2a* and *Q2b*, we inferred the relationships between planning, non-targeted exploration, and problem-solving competency via multidimensional latent variable analysis (Model 2; see Figure 2). That is, we placed the three latent variables together with their correlations at the latent variable level (see the solid arrows between the latent variables in Figure 2) and pairwise residual correlations at the observed variable level (see the dashed arrows between the observed indicators in Figure 2). The covariances between problem-solving competency and planning and between problem-solving competency and non-targeted exploration address *Q2a* and *Q2b* at the latent variable level, respectively. A positive covariance would imply that, generally speaking, planning more or conducting more non-targeted operations is positively related to problem-solving competency. Given the diversity of tasks (e.g., interfaces and complexity), the answers to *Q2a* and *Q2b* might differ between tasks. Hence, we added pairwise residual correlations between the three indicators if they were derived from the same task. For example, for Task 1, we included the residual correlations between P1, E1, and PS1. These residual correlations help explain task-specific relationships among the indicators not captured by the covariances between the latent variables. For example, it could be possible that the overall relationship between non-targeted exploration and problem-solving competency is positive, but for certain tasks exploring more impairs task performance, namely negative task-specific relationships. The specified model is similar to De De Boeck and Scalise's (2019) model, which used time-on-task, the number of actions, and responses as indicators of latent speed, latent action, and latent performance, respectively, in the domain of PISA 2015 collaborative problem-solving. They also considered specific hypotheses about relationships between the residuals of the indicators that were based on the same tasks.

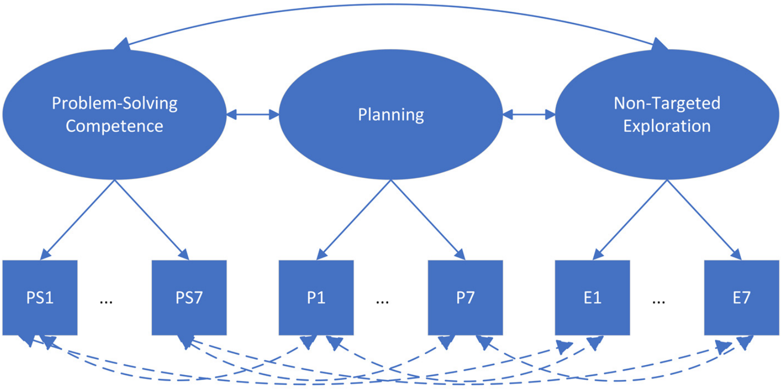


**Figure 2.** An illustration of Model 2. Note. P = planning indicator (i.e., longest duration); PS = task scores; E = non-targeted exploration indicator (i.e., the number of initial non-targeted operations). The numbers 1 to 7 indicate the position of the task in the booklet. Ellipses = latent variables; Rectangles = observed variables. The solid lines with double arrows indicate the covariance between the latent variables. The dashed lines with double arrows indicate the residual correlations between observed indicators.

To estimate the models, we used the lavaan package (Rosseel 2012) in R 4.1.0 (R Core Team 2013) with the diagonally weighted least squares (DWLS) estimator and treated the observed data as ordered categorical variables. Missing values were handled by pairwise deletion. By convention, the means and variances of the latent variables were constrained as zeros and ones for the purpose of model identification, respectively. We evaluated the model fit with a robust chi-square test of fit and used the criteria the root mean square error of approximation (*RMSEA*) and the standardized root mean square residual (*SRMR*). *RMSEA* assesses how far a specified model is away from an ideal model, and *SRMR* evaluates the difference between the residuals of the model-implied covariance matrix and the observed covariance matrix. Hence, the lower *RMSEA* and *SRMR* are, the better the model fit with the data. The cutoff values are 0.06 and 0.08 for *RMSEA* and *SRMR*, respectively (Hu and Bentler 1998).

## 3. Results

We begin this section with a description of the sample characteristics. Among the 1325 participants, the average age was 39 years old (SD = 14), and 53% were female. Around 9%, 40%, and 51% of the participants' highest level of schooling was less than high school, high school, or above high school, respectively. For the employment status, 66% of the participants were employed or self-employed, 3% retired, 8% not working and looking for work, 11% students, 6% doing unpaid household work, and 6% other jobs. PIAAC categorized respondents' performance on the PS-TRE domain in four levels: less than level 1 (19% in the US dataset), level 1 (42% in the US dataset), level 2 (36% in the US dataset), and level 3 (3% in the US dataset). Higher levels indicate better proficiency.

With respect to the responses on the PS-TRE tasks, some omission behaviors were observed for the tasks. There were on average 127 participants (range = [53, 197]) who did not interact with single tasks and requested the next task directly. Figure 3 plots the frequency of the derived indicators after the recoding procedure. The distributions of the planning indicator were almost evenly distributed across the four categories. However, the distributions of the other indicators were somewhat diverse depending on the items. For example, only a small proportion (2.4%) of participants did not try any non-targeted operations in Task 3, but more than one fourth (29%) did not explore Task 7.

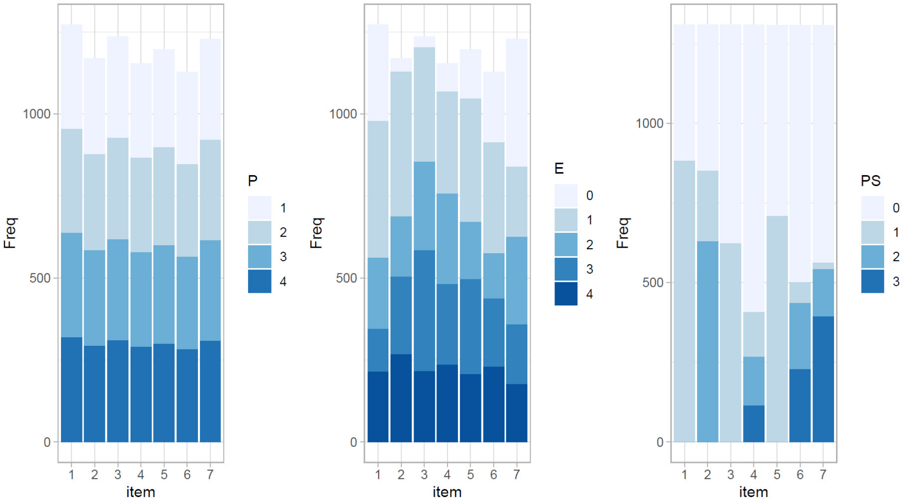


**Figure 3.** The frequency plot of planning (P), non-targeted exploration (E), and problem-solving competency (PS) indicators. The longest duration could not be zero, so the categories of the planning indicator consisted of only four values.

Next, we present the results relevant to *Q1a*, *Q1b*, and *Q1c* based on the single-factor CFA models for planning (Model 1a), non-targeted exploration (Model 1b), and problem-solving competency (Model 1c). Table 2 presents the model fit indices and the standardized results for factor models. For the planning measurement model, although the robust chi-square test was significant ($p$ = .013), the model fit indices (*RMSEA* = 0.021 (*se* = 0.006); *SRMR* = 0.042 (*se* = 0.003)) were lower than the cutoff values 0.06 and 0.08 (Hu and Bentler 1998), thus indicating good approximate model fit. All the factor loadings in Model 1a were significant, ranging from 0.491 to 0.691. The higher factor loading indicates a stronger relationship between the indicator and the latent variable, and thus the latent variable can account for more of the variability of the indicator. The results for the model fit and factor loadings provided evidence of validity for the construct planning. This conclusion also applied to the measurement model (Model 1c) for problem-solving competency (*RMSEA* < 0.001 (*se* = 0.003); *SRMR* = 0.020 (*se* = 0.003); nonsignificant chi-square test, $p$ = .901). The factor loadings ranged from 0.636 to 0.813. For the non-targeted exploration measurement model (Model 1b), the model fit indices (*RMSEA* = 0.014 (*se* = 0.007); *SRMR* = 0.044 (*se* = 0.004)) were satisfactory, and the robust chi-square test was nonsignificant ($p$ = .134). However, the factor loadings varied a lot (see Table 2). Tasks 3 and 4 had the highest factor loadings, whereas the last two tasks had the lowest with values less than 0.2. That is, although the non-targeted exploration indicators in PS-TRE2 generally measure the same construct, the impact of the latent non-targeted exploration on the observed indicators differed across tasks.

**Table 2.** Standardized results for the single-factor models.

| Variable | Estimate | SE | *p* |
|---|---|---|---|
| *Model 1a: Robust $\chi^2$ (35) = 56.179 (p = .013), RMSEA = 0.021 (se = 0.006), SRMR = .042 (se = 0.003)* | | | |
| P1 | 0.531 | 0.028 | <.001 |
| P2 | 0.648 | 0.025 | <.001 |
| P3 | 0.691 | 0.022 | <.001 |
| P4 | 0.662 | 0.025 | <.001 |
| P5 | 0.491 | 0.029 | <.001 |
| P6 | 0.639 | 0.027 | <.001 |
| P7 | 0.663 | 0.023 | <.001 |
| *Model 1b: Robust $\chi^2$ (42) = 52.208 (p = .134), RMSEA = 0.014 (se = 0.007), SRMR = .045 (se = 0.004)* | | | |
| E1 | 0.328 | 0.043 | <.001 |
| E2 | 0.264 | 0.045 | <.001 |
| E3 | 0.414 | 0.048 | <.001 |
| E4 | 0.611 | 0.056 | <.001 |
| E5 | 0.298 | 0.043 | <.001 |
| E6 | 0.179 | 0.046 | <.001 |
| E7 | 0.125 | 0.043 | .003 |
| *Model 1c: Robust $\chi^2$ (28) = 18.892 (p = .901), RMSEA < 0.001 (se = 0.003), SRMR = 0.020 (se = 0.003)* | | | |
| PS1 | 0.778 | 0.025 | <.001 |
| PS2 | 0.786 | 0.020 | <.001 |
| PS3 | 0.684 | 0.026 | <.001 |
| PS4 | 0.813 | 0.019 | <.001 |
| PS5 | 0.758 | 0.024 | <.001 |
| PS6 | 0.636 | 0.025 | <.001 |
| PS7 | 0.723 | 0.022 | <.001 |

Note: P = the planning indicator; E = the non-targeted exploration indicator; PS = the problem-solving indicator.

Subsequently, we present the results of Model 2. If we ignored the residual correlations of the indicators (i.e., the task-dependent effect), the model fit indices exceeded the cutoff values (*RMSEA* = 0.071 > 0.06, *se* = 0.002; *SRMR* = 0.096 > 0.08, *se* = 0.002). This suggests that only considering the overall relationships between the latent variables and excluding the task-dependent relationships did not fit well with the data. In Model 2, the residual correlations were included, and the model fit indices (*RMSEA* = 0.055 < 0.06, *se* = 0.002; *SRMR* = 0.077 < 0.08, *se* = 0.002) improved and implied an acceptable goodness-of-fit (Hu and Bentler 1998). Hence, considering the task-specific effects fit the data substantially better. One obvious difference between single measurement models and the full model occurred in the factor loadings of the non-targeted exploration indicators. In the full model, the latent non-targeted exploration could capture only the common features underlying Tasks 3 and 4, whose factor loadings exceeded 0.4.

Regarding the relationship between planning and problem-solving competency (i.e., *Q2a*), we begin by addressing the latent variable levels, namely their overall relationship. The correlation between latent planning and problem-solving competency was −0.093 ($p$ = .007, $se$ = 0.035). That is, the overall effect of planning on problem-solving was negative, but the magnitude of the effect was rather small. This result was similar to Eichmann et al.'s (2019) study, where the longest duration was not related to task success on average. For *Q2a* on the observed data level, namely the task-dependent relationships, Table 3 presents the relevant results that suggested the residual correlations were not negligible. Specifically, half of the residual correlations were positive, and the other half were negative. For Tasks 3, 4, and 5, after controlling for the latent variables in the model, spending more time on planning contributed to task performance, whereas spending more time on planning in Tasks 1, 6, and 7 impaired task performance. That is, the relationships between the longest duration indicator and task scores varied a lot across the tasks.

**Table 3.** Standardized results of the residual correlations in Model 2.

| Variable | Estimate | SE | *p* |
|---|---|---|---|
| PS1 with P1 | −0.374 | 0.037 | <.001 |
| PS2 with P2 | −0.068 | 0.034 | .365 |
| PS3 with P3 | 0.249 | 0.035 | <.001 |
| PS4 with P4 | 0.569 | 0.033 | <.001 |
| PS5 with P5 | 0.609 | 0.034 | <.001 |
| PS6 with P6 | −0.181 | 0.035 | .002 |
| PS7 with P7 | −0.155 | 0.033 | .013 |
| PS1 with E1 | 0.127 | 0.033 | .014 |
| PS2 with E2 | 0.234 | 0.032 | <.001 |
| PS3 with E3 | 0.179 | 0.024 | <.001 |
| PS4 with E4 | 0.066 | 0.030 | .299 |
| PS5 with E5 | 0.044 | 0.034 | .428 |
| PS6 with E6 | −0.796 | 0.025 | <.001 |
| PS7 with E7 | −0.038 | 0.032 | .408 |
| P1 with E1 | −0.076 | 0.033 | .057 |
| P2 with E2 | −0.002 | 0.033 | .973 |
| P3 with E3 | 0.059 | 0.028 | .233 |
| P4 with E4 | 0.240 | 0.031 | <.001 |
| P5 with E5 | 0.220 | 0.031 | <.001 |
| P6 with E6 | 0.120 | 0.034 | .007 |
| P7 with E7 | 0.208 | 0.032 | <.001 |

Note: P = the planning indicator; E = the non-targeted exploration indicator; PS = the problem-solving indicator.

Regarding *Q2b*, as hypothesized, non-targeted exploration showed a strong positive relationship with problem-solving competency with a factor correlation equal to 0.887 ($p$ < .001, $se$ = 0.034). However, the answer to *Q2b* on the observed data level differed across tasks. The residual correlations between the responses and the non-targeted exploration indicators were significant and positive in the first three tasks but negative in Task 6 (see Table 3). That is, after considering the positive relationship between non-targeted exploration and problem-solving competency, different tasks showed distinct impacts on task performance. In addition, the residual correlations between the indicators of planning and non-targeted exploration by and large increased with the positions of the tasks. Engagement might be one explanation for this result. Specifically, participants who kept engaging in the assessment tended to invest more time in planning and more exploratory behaviors than those who gradually lost patience.

## 4. Discussion

In this article, we focused on planning, non-targeted exploration, and problem-solving competency using process measures and task performance in the 2012 PIAAC PS-TRE domain. We assessed the internal construct validity of the derived indicators and investigated their relationships using multidimensional latent variable analysis.

### *4.1. Summary of the Study*

Our results provide additional evidence for the internal construct validity of the indicators of planning and problem-solving competency. It suggested that the latent planning greatly captured the common variance of the longest duration indicators and was relatively stable across tasks. However, the CFA results indicated that latent non-targeted exploration exerted varied influences on different tasks. The task interfaces can provide a potential explanation for the result. If the interfaces such as spreadsheets or emails contained features that are commonly used by respondents, it would likely be less necessary to explore these buttons to acquire new information. In contrast, novel information was embedded in a web environment in Tasks 3, 4, and 7, requiring potentially more non-targeted exploration, while Task 7 provided extra hints for non-necessary operations and thus prevented some non-targeted behaviors. In short, the familiarity of the presented environments and hints might weaken the influence of the latent non-targeted exploration.

After interpreting the internal construct validity of the process indicators, we then interpret the task-dependent relationships between planning and problem-solving competency. Task difficulty was not critical in explaining the diverse relationships after we inspected the task difficulty for each item provided by PIAAC (OECD 2013), a finding that was in line with Eichmann et al. (2019) who used the PISA 2012 problem-solving tasks. Instead, more specific task features can provide some insights. If some tasks (e.g., Task 4) require respondents to integrate complex information, investing more time in planning helps problem-solving (Mumford et al. 2001). Moreover, the relevance of information also mattered. Being stuck with irrelevant information can lead to biased planning (Mumford et al. 2001). For instance, we found that unsuccessful respondents tended to spend the longest duration on irrelevant emails compared with successful respondents in Task 6.

The other research interest of the study is the relationships between problem-solving competency and non-targeted exploration. The positive overall relationship between non-targeted exploration and problem-solving competency on the latent trait level indicated that non-targeted exploration facilitated representing and further contributed to successful task completion (Dormann and Frese 1994; Kapur 2008). However, the negative residual correlation for Task 6 implied that exploring too much was detrimental to solving the task. Paying too much attention to irrelevant information might complicate the problem and result in cognitive overload (Frese and Keith 2015). A common pattern for successful problem-solving involved actively trying some non-targeted operations or goal-directed behaviors to expand the problem space, distinguishing the features of these operations, and focusing on goal-directed behaviors to reach the desired state.

### *4.2. Contributions and Limitations*

This article offers several contributions. From a theoretical perspective, we examined the internal construct validity of process indicators across multiple tasks, whereas many relevant studies have been limited to single items (e.g., Ulitzsch et al. 2021). Combining data from multiple tasks utilizes the information from the assessment to a greater extent and potentially provides more evidence for the stability of the conclusions. We found that the process indicators differed in the extent of internal construct validity, which suggested that researchers should carefully consider applying the measures from one task to another task even though both tasks are designed to measure the same concept. For practitioners, the longest duration can be employed as a good indicator for planning in other information-processing problems similar to the PS-TRE tasks, whereas non-targeted exploration would be less suitable to apply to routine problems with little novel information. On the contrary, if the task is rich in new information that respondents can explore to acquire, the amount of non-targeted exploration would be able to capture the common pattern of exploratory behaviors.

Regarding the research topics, our results provide evidence for the functions of planning and non-targeted exploration in problem-solving based on human–computer interactions, deepening the understanding of their relationships in dynamic problems. The insight into the processes of complex problem-solving is crucial for educational systems since one important mission of education is to prepare students to become better problem-solvers (OECD 2014). Our results can potentially facilitate educational practice aiming at improving problem-solving skills. For example, it would be promising to implement a computer-simulated agent to help problem-solvers in terms of planning and non-targeted exploration. Specifically, if an individual has spent a long time planning in a dynamic problem without interacting with the task environment, the agent can offer a hint to encourage exploratory behaviors if the individual is not familiar with the task environment. In another circumstance, if an individual engaged in too much non-targeted exploration rapidly, the agent can advise spending more time on planning a strategy when the task requires respondents to incorporate complex information. Besides the development of digital tools, test developers can also compare the relationships between planning, non-targeted exploration, and task performance with the desired design to reflect on the task design. For example, if a task is designed to benefit from planning, the relationship between the longest duration and task performance should be positive; otherwise, test developers would need to reconsider their design.

Some limitations of this study should also be noted. First, the indicator of non-targeted exploration requires researchers to define goal-directed and non-targeted operations that can be difficult for some types of problems. Second, the longest duration indicator reflects only the quantity of the planning, which does not necessarily imply the quality of the planning. Future studies can assess the quality of plans in dynamic problems and examine their relationship with task performance. In addition, similar to Eichmann et al. (2019), our definition of planning is broad in nature. Although we excluded the durations at the end of the tasks (e.g., reflecting process) in identifying the planning process, the longest duration can actually refer to the monitoring process. Third, although our indicators were based on previous studies, the underlying meaning of the latent variables must be interpreted carefully. Fourth, the current data are from the 2012 PIAAC PS-TRE domain, the core of which is information-processing skills (Greiff et al. 2017). However, other international assessments have various focuses, which may show different relationships between planning, non-targeted exploration, and problem-solving competency.

## 5. Conclusions

This study derived process indicators of planning and non-targeted exploration from the existing literature (Eichmann et al. 2019, 2020a, 2020b). Our results provide evidence for the internal construct validity of the planning indicator and response scores across multiple PS-TRE items, whereas the non-targeted exploration indicator was more challenging to be analyzed simultaneously across tasks when considering the dependency of the indicators from the same item. In addition, non-targeted exploration had a strong positive relationship with problem-solving competency. The results of residual correlations provided more detailed and diverse relationships between task performance, planning, and non-targeted exploration on the task level.

**Author Contributions:** Conceptualization, M.Z.; methodology, M.Z. and B.A.; validation, M.Z., B.A. and S.G.; formal analysis, M.Z.; writing—original draft preparation, M.Z.; writing—review and editing, B.A. and S.G.; visualization, M.Z.; supervision, B.A. and S.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Ethical review and approval were waived for this study due to the anonymized data and secondary analysis in this study.

**Informed Consent Statement:** The data used in this article are anonymized secondary data and neither consent to participate or consent for publication nor ethics approval were required for the current study.

**Data Availability Statement:** The data used in this article can be found in the GESIS Data Catalogue (https://search.gesis.org/research_data/ZA6712?doi=10.4232/1.12955, accessed on 19 October 2021).

**Acknowledgments:** This paper is dedicated to Sigrid Blömeke, a valued colleague, who recently passed away. We mourn the loss of a wonderful person, a supportive leader, and a stellar researcher. This paper is dedicated to her memory and will always remind us to continue her scientific legacy.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

The recoding rules for log-events:

The log-events were recoded using the following rules:

- We kept only the events implemented by the respondent and deleted the system events triggered by the respondent's interaction event. For instance, when a respondent clicked on the "Add page" button in the bookmark pop-up window, three events were logged with the same timestamps: BOOKMARK_ADD, BUTTON, and DOACTION. In this case, we kept only BOOKMARK_ADD because it was sufficient for describing the operation implemented by the respondent.
- We aggregated the event type KEYPRESS. When a key is pressed, a KEYPRESS event with an ASCII value is logged. Because typing a string (e.g., a name) is regarded as a single operation, we aggregated consecutive KEYPRESS events as a single KEYPRESS event.
- All events from a combo-box (e.g., a SORT pop-up window) with several sorting rules were aggregated according to the final state of the SORT window.

## Appendix B

**Table A1.** Descriptive statistics for the raw process indicators without transformations.

| Raw Indicator | Mean | SD | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| P1 | 66.75 | 59.85 | 1.09 | 1149 | 7.15 | 101.14 |
| P2 | 55.63 | 56.10 | 4.46 | 1317 | 11.51 | 227.32 |
| P3 | 41.04 | 29.33 | 1.91 | 432 | 4.67 | 42.70 |
| P4 | 43.32 | 78.19 | 6.12 | 2421 | 24.79 | 739.48 |
| P5 | 48.58 | 33.65 | 1.85 | 313 | 2.41 | 9.84 |
| P6 | 34.99 | 322.84 | 4.25 | 10847 | 33.30 | 1112.34 |
| P7 | 27.88 | 65.25 | 3.66 | 2157 | 28.70 | 921.94 |
| E1 | 1.94 | 2.28 | 0 | 38 | 4.61 | 52.76 |
| E2 | 8.50 | 18.90 | 0 | 204 | 5.82 | 44.57 |
| E3 | 7.44 | 3.98 | 0 | 17 | .07 | −1.11 |
| E4 | 6.48 | 5.15 | 0 | 35 | 1.08 | 1.49 |
| E5 | 3.81 | 3.56 | 0 | 30 | 1.91 | 6.51 |
| E6 | 8.37 | 10.68 | 0 | 65 | 1.27 | .63 |
| E7 | 3.6 | 5.56 | 0 | 46 | 3.21 | 13.26 |

Note: P = the planning indicator. E = the non-targeted exploration indicator.

## References

AERA. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Albert, Dustin, and Laurence Steinberg. 2011. Age differences in strategic planning as indexed by the Tower of London. *Child Development* 82: 1501–17. [CrossRef]

Arlin, Patricia Kennedy. 1989. The problem of the problem. In *Everyday Problem Solving: Theory and Applications*. Edited by Jan D. Sinnott. New York: Wittenborn, pp. 229–37.

Bartholomew, David J., Martin Knott, and Irini Moustaki. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. Hoboken: John Wiley & Sons.

Bell, Bradford S., and Steve W. J. Kozlowski. 2008. Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology* 93: 296–316. [CrossRef]

Berbeglia, Gerardo, Jean-François Cordeau, Irina Gribkovskaia, and Gilbert Laporte. 2007. Static pickup and delivery problems: A classification scheme and survey. *TOP* 15: 1–31. [CrossRef]

Box, George E., and David R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* 26: 211–43. [CrossRef]

Brown, Timothy A. 2015. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford publications.

Bruning, Roger H., Gregory J. Schraw, Monica M. Norby, and Royce R. Ronning. 2004. *Cognitive Psychology and Instruction*, 4th ed. Upper Saddle River: Merrill Prentice Hall.

Chen, Yunxiao, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. 2019. Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology* 10: 486. [CrossRef] [PubMed]

Chung, Ji Eun, and Stuart Elliott. 2015. *Adults, Computers and Problem Solving: "What's the Problem?" OECD Skills Studies*. Paris: OECD Publishing.

Csapó, Benő, and Joachim Funke. 2017. *The Nature of Problem Solving*. Paris: OECD Publishing.

De Boeck, Paul, and Kathleen Scalise. 2019. Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology* 10: 1280. [CrossRef]

Dormann, Tanja, and Michael Frese. 1994. Error training: Replication and the function of exploratory behavior. *International Journal of Human-Computer Interaction* 6: 365–72. [CrossRef]

Eichmann, Beate, Frank Goldhammer, Samuel Greiff, Liene Brandhuber, and Johannes Naumann. 2020a. Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology* 112: 1546–62. [CrossRef]

Eichmann, Beate, Frank Goldhammer, Samuel Greiff, Liene Pucite, and Johannes Naumann. 2019. The role of planning in complex problem solving. *Computers & Education* 128: 1–12. [CrossRef]

Eichmann, Beate, Samuel Greiff, Johannes Naumann, Liene Brandhuber, and Frank Goldhammer. 2020b. Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning* 36: 933–56. [CrossRef]

Frese, Michael, and Nina Keith. 2015. Action errors, error management, and learning in organizations. *Annual Review of Psychology* 66: 661–87. [CrossRef] [PubMed]

Frese, Michael, Felix Brodbeck, Torsten Heinbokel, Christina Mooser, Erik Schleiffenbaum, and Petra Thiemann. 1991. Errors in training computer skills: On the positive function of errors. *Human-Computer Interaction* 6: 77–93. [CrossRef]

Greiff, Samuel, Katharina Scheiter, Ronny Scherer, Francesca Borgonovi, Ann Britt, Art Graesser, Muneo Kitajima, and Jean-François Rouet. 2017. *Adaptive Problem Solving: Moving towards a New Assessment Domain in the Second Cycle of PIAAC*. OECD Education Working Papers 156. Paris: OECD Publishing.

Greiff, Samuel, Sascha Wüstenberg, and Francesco Avvisati. 2015. Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education* 91: 92–105. [CrossRef]

Greiff, Samuel, Sascha Wüstenberg, Benő Csapó, Andreas Demetriou, Jarkko Hautamäki, Arthur C. Graesser, and Romain Martin. 2014. Domain-general problem solving skills and education in the 21st century. *Educational Research Review* 13: 74–83. [CrossRef]

Greiff, Samuel, Sascha Wüstenberg, Daniel V. Holt, Frank Goldhammer, and Joachim Funke. 2013. Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development* 61: 407–21. [CrossRef]

Hayes-Roth, Barbara, and Frederick Hayes-Roth. 1979. A cognitive model of planning. *Cognitive Science* 3: 275–310. [CrossRef]

He, Qiwei, and Matthias von Davier. 2016. Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In *Handbook of Research on Technology Tools for Real-World Skill Development*. Hershey: IGI Global, pp. 750–77.

He, Qiwei, Francesca, Borgonovi, and Marco Paccagnella. 2021. Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education* 166: 104170. [CrossRef]

Hu, Li-tze, and Peter M. Bentler. 1998. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods* 3: 424–53. [CrossRef]

Jonassen, David H. 2000. Toward a design theory of problem solving. *Educational Technology Research and Development* 48: 63–85. [CrossRef]

Jöreskog, Karl Gustav. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34: 183–202. [CrossRef]

Kapur, Manu. 2008. Productive failure. *Cognition and Instruction* 26: 379–424. [CrossRef]

Keith, Nina, and Michael Frese. 2005. Self-regulation in error management training: Emotion control and metacognition as mediators of performance effects. *Journal of Applied Psychology* 90: 677–91. [CrossRef] [PubMed]

Mayer, Richard E., and Merlin C. Wittrock. 2006. Problem solving. In *Handbook of Educational Psychology*, 2nd ed. Edited by Patricia A. Alexander and Philip H. Winne. Mahwah: Erlbaum, pp. 287–303.

Mumford, Michael D., Rosemary A. Schultz, and Judy R. Van Doorn. 2001. Performance in planning: Processes, requirements, and errors. *Review of General Psychology* 5: 213–40. [CrossRef]

OECD. 2012. *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*. Paris: OECD Publishing.

OECD. 2013. *Technical Report of the Survey of Adult Skills (PIAAC)*. Paris: OECD Publishing.

OECD. 2014. *PISA 2012 Results: Creative Problem Solving: Students' Skills in Tackling Real-Life Problems (Volume V)*. Paris: OECD Publishing.

OECD. 2017. *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*. Cologne: GESIS Data Archive.

Peterson, Ryan A., and Joseph E. Cavanaugh. 2019. Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics* 47: 2312–27. [CrossRef] [PubMed]

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Ridgway, Jim, and Sean McCusker. 2003. Using computers to assess new educational goals. *Assessment in Education: Principles, Policy & Practice* 10: 309–328. [CrossRef]

Rosseel, Yves. 2012. Lavaan: An r Package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software* 48: 1–36. [CrossRef]

Schunk, Dale H. 2003. Self-regulation and learning. In *Handbook of Psychology*. Edited by William M. Reynolds and Gloria E. Miller. New York: Wiley, vol. 7, pp. 59–78.

Stadler, Matthias, Christoph Niepel, and Samuel Greiff. 2019. Differentiating between static and complex problems: A theoretical framework and its empirical validation. *Intelligence* 72: 1–12. [CrossRef]

Ulitzsch, Esther, Qiwei He, Vincent Ulitzsch, Hendrik Molter, André Nichterlein, Rolf Niedermeier, and Steffi Pohl. 2021. Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika* 86: 190–214. [CrossRef]

Unterrainer, Josef M., and Adrian M. Owen. 2006. Planning and problem solving: From neuropsychology to functional neuroimaging. *Journal of Physiology-Pari* 99: 308–17. [CrossRef]

Unterrainer, Josef M., Benjamin Rahm, Rainer Leonhart, Christian C. Ruff, and Ulrike Halsband. 2003. The Tower of London: The impact of instructions, cueing, and learning on planning abilities. *Cognitive Brain Research* 17: 675–83. [CrossRef]

Wüstenberg, Sascha, Samuel Greiff, and Joachim Funke. 2012. Complex problem solving—More than reasoning? *Intelligence* 40: 1–14. [CrossRef]

Xiao, Yan, Paul Milgram, and Daniel John Doyle. 1997. Planning behavior and its functional role in interactions with complex systems. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 27: 313–24. [CrossRef]

Yuan, Jianlin, Yue Xiao, and Hongyun Liu. 2019. Assessment of collaborative problem solving based on process stream data: A new paradigm for extracting indicators and modeling dyad data. *Frontiers in Psychology* 10: 369. [CrossRef] [PubMed]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Paper III

# Fast estimation of multiple group generalized linear latent variable models for categorical observed variables

**Björn Andersson, Shaobo Jin, Maoxin Zhang**

Published in *Computational Statistics & Data Analysis*.

Computational Statistics and Data Analysis 182 (2023) 107710

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Fast estimation of multiple group generalized linear latent variable models for categorical observed variables

Björn Andersson[a,*], Shaobo Jin[b,c], Maoxin Zhang[a]

[a] Centre for Educational Measurement at the University of Oslo (CEMO), P.O. Box 1161 Forskningsparken, Oslo, 0318, Norway
[b] Department of Statistics, Uppsala University, Box 513, Uppsala, 751 20, Sweden
[c] Department of Mathematics, Uppsala University, Box 480, Uppsala, 751 06, Sweden

ARTICLE INFO

*Article history:*
Received 28 April 2022
Received in revised form 20 October 2022
Accepted 28 January 2023
Available online 2 February 2023

*Keywords:*
Latent variable models
Item response theory
Integral approximation
Gauss-Hermite quadrature
Laplace approximation

ABSTRACT

A computationally efficient method for marginal maximum likelihood estimation of multiple group generalized linear latent variable models for categorical data is introduced. The approach utilizes second-order Laplace approximations of the integrals in the likelihood function. It is demonstrated how second-order Laplace approximations can be utilized highly efficiently for generalized linear latent variable models by considering symmetries that exist for many types of model structures. In a simulation with binary observed variables and four correlated latent variables in four groups, the method has similar bias and mean squared error compared to adaptive Gauss-Hermite quadrature with five quadrature points while substantially improving computational efficiency. An empirical example from a large-scale educational assessment illustrates the accuracy and computational efficiency of the method when compared against adaptive Gauss-Hermite quadrature with three, five, and 13 quadrature points.

## 1. Introduction

When estimating latent variable models for categorical observed variables, such as generalized linear latent variable models or item response theory models, marginal maximum likelihood estimation is typically used. With marginal maximum likelihood, integrals without an explicit solution must be calculated. Standard estimation methods are based on Gauss-Hermite quadrature approximations (Bock and Aitkin, 1981), which are highly efficient for models with one or two latent variables but quickly decrease in efficiency with higher-dimensional problems. Adaptive quadrature (Schilling and Bock, 2005; Cagnone and Monari, 2013), which concentrates the region of integration to the most relevant part for each integral, is a partial solution but with more than four dimensions the computational expense means that adaptive quadrature becomes impractical if high accuracy is desired. Another approach is to use a simulation-based method such as the Metropolis-Hastings Robbins-Monro method (Cai, 2010a) or a Monte Carlo method (Zhu et al., 2005). However, the simulation-based methods can be slow to converge to a local maximum with small sample sizes and ensuring that proper convergence has been attained is often challenging and time-consuming. Efficient approximation methods such as the variational approximation have also been proposed (Hui et al., 2017; Niku et al., 2019; Cho et al., 2021), but these methods perform relatively poorly with few observed variables and have not been implemented for many types of models.

[*] Corresponding author.
*E-mail address:* bjorn.andersson@cemo.uio.no (B. Andersson).

https://doi.org/10.1016/j.csda.2023.107710

Another approach is to use a first-order Laplace approximation to approximate the required integrals (Huber et al., 2004). However, estimation methods based on first-order Laplace approximations often have convergence problems and high bias (Joe, 2008), especially with binary data or complex models (Andersson and Xin, 2021). To remedy this, second-order Laplace approximations of the log-likelihood (Shun, 1997; Thomas, 1993; Bianconcini, 2014; Raudenbush et al., 2000) or second-order Laplace approximations of the gradient of the log-likelihood (Bianconcini and Cagnone, 2012) have been used. Such higher-order Laplace approximations have shown promise in providing a computationally efficient yet accurate estimation method for models with many latent variables, such as in Andersson and Xin (2021) where up to 12 correlated latent variables were used with an independent-clusters structure and ordinal observed variables. In contrast to this, some previous studies found that the second-order Laplace approximation was computationally much less efficient compared to adaptive quadrature with 5 quadrature points in each dimension for exploratory factor analysis with ordinal data with up to four dimensions (Bianconcini, 2014). As we will show in the present study, the efficiency of higher-order Laplace approximations is highly dependent on the structure of the model used and implementations that disregard the structure will be inefficient for most types of models. However, if the structure is exploited in the implementation of the second-order Laplace approximation, substantial computational gains can be obtained which makes the method highly computationally efficient for many types of models when compared to adaptive quadrature approximations which have the same theoretical approximation accuracy.

A hindrance to methods which use second-order Laplace approximations is that they require the derivation of higher-order derivatives which depend on the type of measurement model specified. This makes higher-order Laplace approximations difficult to efficiently implement and generalize across different types of models. Hence, so far, the available implementations of second-order Laplace approximations of the log-likelihood which are relevant to latent variable models for categorical data have been limited to generalized linear models (Raudenbush et al., 2000), generalized linear mixed models (Noh and Lee, 2007), confirmatory factor analysis models for ordinal data (Bianconcini, 2014; Jin et al., 2018), independent-clusters item response theory models (Thomas, 1993; Andersson and Xin, 2021) and nonlinear structural equation models (Jin et al., 2020). Thus, additional research is needed to implement estimation of multiple group generalized linear latent variable models with second-order Laplace approximations and to investigate its estimation properties.

The objective of the current work is then to develop a computationally efficient estimation method based on a second-order Laplace approximation to estimate multidimensional generalized linear latent variable models with categorical observed variables, with support for an arbitrary model structure and with multiple groups. There are three main contributions of the present study in relation to the existing literature. First, we derive an estimation algorithm that uses a second-order Laplace approximation to the marginal log-likelihood function for generalized linear latent variable models for categorical observed variables which supports a general model structure. Here, we also detail how the second-order Laplace approximation can be highly efficiently implemented by accounting for the structure of the particular model used. Second, we implement the second-order Laplace approximation estimation method for multiple group models where parameter invariance between groups can be established and where the mean vectors and covariance matrices of the latent variable in multiple groups can be estimated. Third, we compare the second-order Laplace approximation method to an implementation of adaptive Gauss-Hermite quadrature, that uses the same underlying code base as the second-order Laplace method, in terms of the estimation accuracy and precision and in terms of the computational efficiency.

The paper is structured as follows. We first introduce the modeling framework used and then present the second-order Laplace approximation estimation method along with a discussion of some of its properties with models commonly used in applied measurement. Then, based on a simulation study, we contrast and compare the proposed approach to adaptive Gauss-Hermite quadrature and discuss the advantages and disadvantages of the method based on theoretical and practical considerations. Subsequently, an empirical example from an international large-scale assessment is used to illustrate the application of the Laplace approximations and adaptive quadrature methods. Lastly, we discuss our findings and provide recommendations for applied work.

## 2. Methods

### 2.1. Models

With latent variable models for categorical data, we model the response probabilities for each category of a set of discrete observed variables $i \in \{1, \ldots, I\}$ conditional on a latent variable. Define $P_{ic}(\boldsymbol{z})$ as the probability, conditional on the $p \times 1$ latent variable vector $\boldsymbol{z}$, to observe category $c$ of observed variable $Y_i$ which has $m_i$ possible outcomes. We assume conditional independence such that the joint probability for multiple random variables $Y_1 = y_1, \ldots, Y_I = y_I$, conditional on $\boldsymbol{z}$, can be factorized as

$$P(Y_1 = y_1, \ldots, Y_I = y_I|\boldsymbol{z}) = \prod_{i=1}^{I} P_{iy_i}(\boldsymbol{z}), \tag{1}$$

where the individual $P_{iy_i}$ can be based on, for example, confirmatory factor analysis with categorical data, the generalized partial credit model (Muraki, 1992), the graded response model (Samejima, 1969), or the nominal response model

(Bock, 1972). These three specific models are all types of generalized linear latent variable models (Huber et al., 2004) and also fall within the framework of generalized linear latent and mixed models (Rabe-Hesketh et al., 2004). Let $\boldsymbol{b}_i$ be a $m_i \times 1$ vector of intercept parameters, with entries $b_{ic}$ such that $b_{i1} = 0$. For the graded response model, with a $p \times 1$ vector $\boldsymbol{a}_i$ of slope parameters, we have

$$P_{ic}(\boldsymbol{z}) = P^*_{ic}(\boldsymbol{z}) - P^*_{i(c+1)}(\boldsymbol{z}), \tag{2}$$

where

$$P^*_{ic}(\boldsymbol{z}) = \frac{1}{1 + \exp(-\boldsymbol{a}'_i \boldsymbol{z} - b_{ic})}, \tag{3}$$

with $P^*_{i1}(\boldsymbol{z}) = 1$ and $P^*_{i(m_i+1)}(\boldsymbol{z}) = 0$. For the nominal response model, with a $p \times 1$ vector $\boldsymbol{a}_{ic}$ of slope parameters for each category $c$ such that $\boldsymbol{a}_{i1} = \boldsymbol{0}$, we have

$$P_{ic}(\boldsymbol{z}) = \frac{\exp\left(\boldsymbol{a}'_{ic}\boldsymbol{z} + b_{ic}\right)}{\sum_{c'=1}^{m_i} \exp\left(\boldsymbol{a}'_{ic'}\boldsymbol{z} + b_{ic'}\right)}, \tag{4}$$

and for the generalized partial credit model, a special case of the nominal response model, we have

$$P_{ic}(\boldsymbol{z}) = \frac{\exp\left[\sum_{v=1}^{c}(\boldsymbol{a}'_i\boldsymbol{z} + b_{iv})\right]}{\sum_{c'=1}^{m_i} \exp\left[\sum_{v=1}^{c'}(\boldsymbol{a}'_i\boldsymbol{z} + b_{iv})\right]}. \tag{5}$$

In principle any probability model that satisfies the conditional independence assumption can be used. Let $N_g$ denote the sample size in group $g$ and define $\boldsymbol{y}_{fg}$ as the $I \times 1$ vector of observed variables for an individual $f \in \{1, \ldots, N_g\}$ in group $g \in \{1, \ldots, G\}$ and let $N = \sum_{g=1}^{G} N_g$. The marginal log-likelihood for an individual $f$ in group $g$ is equal to

$$l_{fg}(\boldsymbol{\theta}_g | \boldsymbol{y}_{fg}) = \log \int P(\boldsymbol{y}_{fg}|\boldsymbol{z})\phi(\boldsymbol{z}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)d\boldsymbol{z}, \tag{6}$$

where $\boldsymbol{\theta}_g$ are the unknown parameters in group $g$ and $\phi$ is the multivariate normal density function with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. Define $\boldsymbol{\theta}$ as the vector of all free parameters of the model across all groups. With multiple group models it is possible to evaluate measurement invariance across groups and estimate the mean vectors and covariance matrices in the groups, provided there exist some observed indicators which exhibit invariance (Muthen and Lehman, 1985). Typically, we cannot solve the integral in Equation (6) analytically and it must be approximated.

### 2.2. Likelihood approximation

We consider approximating the marginal log-likelihood with either a second-order Laplace approximation or adaptive Gauss-Hermite quadrature. Here, we present these two methods and outline their properties in terms of accuracy and computational efficiency.

#### 2.2.1. A second-order Laplace approximation

We propose to approximate the integrals in the likelihood function with a second-order Laplace approximation (Shun, 1997) and implement an estimation method based on such approximations. Define the function $h_{fg}(\boldsymbol{z}) = -\log P(\boldsymbol{y}_{fg}|\boldsymbol{z})\phi(\boldsymbol{z}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. The second-order Laplace approximation is based on 1) the estimation of the posterior mode of the latent variable vector for each individual, and 2) derivatives of $h_{fg}$ with respect to $\boldsymbol{z}$ up to the fourth order. Let $\hat{\boldsymbol{z}}_{fg}$ be the posterior mode, equal to the minimizer of $h_{fg}(\boldsymbol{z})$. Define $\hat{h} = h_{fg}(\hat{\boldsymbol{z}}_{fg})$, $\boldsymbol{H}_{fg} = \frac{\partial^2 \hat{h}}{\partial \boldsymbol{z} \partial \boldsymbol{z}'}$ and let $\hat{\boldsymbol{Z}}$ denote the $p \times N$ matrix of posterior modes. We then obtain the second-order Laplace approximation to the log-likelihood as Shun (1997)

$$l^{\text{Lap2}}_{fg}(\boldsymbol{\theta}_g | \boldsymbol{y}_{fg}) = \frac{p}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{H}_{fg}\right| - \hat{h} + \log(1 + \epsilon_{fg}), \tag{7}$$

where, with $b_{jk}$ denoting the $j$-th row entry of the $k$-th column in $\boldsymbol{H}^{-1}_{fg}$,

$$\epsilon_{fg} = -\frac{1}{2}\left[\frac{1}{4}\sum_{jklm}^{p} \frac{\partial^4 \hat{h}}{\partial z_j \partial z_k \partial z_l \partial z_m} b_{jl} b_{km} - \frac{1}{4}\sum_{jklrst}^{p} \frac{\partial^3 \hat{h}}{\partial z_j \partial z_k \partial z_l}\frac{\partial^3 \hat{h}}{\partial z_r \partial z_s \partial z_t} b_{jr} b_{kl} b_{st} \right.$$
$$\left. - \frac{1}{6}\sum_{jklrst}^{p} \frac{\partial^3 \hat{h}}{\partial z_j \partial z_k \partial z_l}\frac{\partial^3 \hat{h}}{\partial z_r \partial z_s \partial z_t}\frac{1}{6} b_{jr} b_{ks} b_{lt}\right]. \tag{8}$$

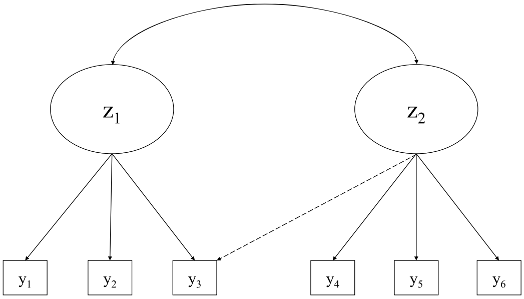


**Fig. 1.** Illustration of a model with six observed variables and two latent variables.

Denote the sums of Equation (8) as sum A, B and C, for the first, second and third entry, respectively. Without considering the model structure, Equation (8) requires the computation of $p^4 + p^6 + p^6$ entries which quickly becomes computationally demanding as the number of latent variables $p$ increases. However, with models structured in a particular way, many of the terms needed for computing the entries in Equation (8) will be zero or repeated. For example, if all observed variables are each related to just one out of many latent variables, the expression reduces to a simple sum and a two-fold sum instead of the four-fold and six-fold sums in Equation (8) (Andersson and Xin, 2021; Noh and Lee, 2007). In our implementation of the second-order Laplace approximation, we avoid computing the same entries multiple times and identify the unique entries, which are products of multiple terms, in each of the sums in Equation (8). Such a procedure was also suggested in the supplementary material of Jin and Andersson (2020). We accomplish this with a computer algorithm at the first step of the estimation process which identifies the entries that must be computed and thus filters out repeated and zero entries. We subsequently weight the unique entries in accordance with the frequency of each entry in the sum.

In addition to avoiding zero and repeated entries in Equation (8), we exploit the expression of the function $h_{fg}(\boldsymbol{z})$ to gain further computational advantages. To illustrate this, first observe that

$$h_{fg}(\boldsymbol{z}) = -\sum_{i=1}^{I} \log P(y_{ifg}|\boldsymbol{z}; \boldsymbol{\alpha}_{ig}) - \log \phi(\boldsymbol{z}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{9}$$

where $y_{ifg}$ denotes the $i$th observed variable for individual $f$ in group $g$ and where $\boldsymbol{\alpha}_{ig}$ is the parameter vector for the $i$th observed variable in group $g$. Define $h_{ifg}(\boldsymbol{z}) = -\log P(y_{ifg}|\boldsymbol{z}; \boldsymbol{\alpha}_{ig})$. Since derivatives of $\log \phi(\boldsymbol{z}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ with respect to $\boldsymbol{z}$ of order three and higher are all zero, we have

$$\frac{\partial^u h_{fg}(\boldsymbol{z})}{\partial z_j \partial z_k \ldots \partial z_v} = \sum_{i=1}^{I} \frac{\partial^u h_{ifg}(\boldsymbol{z})}{\partial z_j \partial z_k \ldots \partial z_v}, \tag{10}$$

for $u > 2$. The implication of expressing the higher-order derivatives in this manner is that, while the higher-order derivative terms in the approximation of the likelihood in Equation (8) may not be zero or equal for a combination of $j, k, \ldots, v$, in many cases a term $\frac{\partial^u h_{ifg}(\boldsymbol{z})}{\partial z_j \partial z_k \ldots \partial z_v}$ for a single observed variable is indeed zero or equal for this combination of $j, k, \ldots, v$. For example, consider the model with six observed variables and two latent variables displayed in Fig. 1 where the ellipses represent the latent variables and the rectangles represent the observed variables. For this model, the unique third derivatives of $h_{fg}$ are $\frac{\partial^3 h_{fg}(\boldsymbol{z})}{\partial z_1^3}$, $\frac{\partial^3 h_{fg}(\boldsymbol{z})}{\partial z_1^2 \partial z_2}$, $\frac{\partial^3 h_{fg}(\boldsymbol{z})}{\partial z_1 \partial z_2^2}$, and $\frac{\partial^3 h_{fg}(\boldsymbol{z})}{\partial z_2^3}$. However, for all observed variables except $y_3$, the only non-zero third-derivatives are $\frac{\partial^3 h_{ifg}(\boldsymbol{z})}{\partial z_1^3}$, for $i \in \{1, 2\}$, and $\frac{\partial^3 h_{ifg}(\boldsymbol{z})}{\partial z_2^3}$, for $i \in \{4, 5, 6\}$. We thus also account for these patterns when computing the entries in the main approximation, beyond accounting for the symmetries that exist for the entries in the two four-fold and six-fold sums in Equation (8).

The filtering process described above means that the second-order Laplace approximation can be implemented with substantial efficiency gains compared to using, for example, adaptive Gauss-Hermite quadrature with the same order of accuracy. Note that, unlike in Shun (1997), we do not remove terms from the approximation in Equation (8) to improve the computational efficiency. Rather, we account for zero entries and symmetries that exist for the models we use.

### *2.2.2. Adaptive Gauss-Hermite quadrature*

Let $\boldsymbol{\Gamma}_{fg}$ be the Cholesky decomposition of the matrix $\boldsymbol{H}_{fg}^{-1}$. The adaptive quadrature approximation to the log-likelihood is then (Jin and Andersson, 2020)

$$l_{fg}^{\mathrm{AGHQ}}(\boldsymbol{\theta}_g|\boldsymbol{y}_{fg}) = \frac{p}{2}\log(2) - \frac{1}{2}\log\left|\boldsymbol{H}_{fg}\right| + \log \sum_{j_1,\ldots,j_p}^{Q} \left[\prod_{k=1}^{p} w_{j_k} \exp\left(q_{j_k}^2\right)\right] \exp\left(h_{fg}(\boldsymbol{z})\Big|^{\boldsymbol{z}=\sqrt{2}\boldsymbol{\Gamma}_{fg}\boldsymbol{q}_{j_1,\ldots,j_p}+\hat{\boldsymbol{z}}_{fg}}\right), \tag{11}$$

where $Q$ denotes the number of quadrature points per dimension, $q_{j_k}$ is the $j_k$-th Gauss-Hermite quadrature point with weight $w_{j_k}$ and $\boldsymbol{q}_{j_1,\ldots,j_p} = (q_{j_1}, \ldots, q_{j_p})'$. The theoretical approximation accuracy of adaptive Gauss-Hermite quadrature depends on the number of quadrature points and the error rate is given by $O\left(I^{-\lfloor (Q+2)/3 \rfloor}\right)$ (Jin and Andersson, 2020), implying that using four to six quadrature points has the same theoretical accuracy as the second-order Laplace approximation.

Adaptive Gauss-Hermite quadrature requires $Q^p$ number of quadrature points, meaning that higher-dimensional models quickly become very computationally demanding to estimate. For example, a four-dimensional model requires a total of 81, 625, and 2401 quadrature points for $Q = 3, 5,$ and 7, respectively. Unlike for the number of entries required with the second-order Laplace approximation, the total number of quadrature points needed for a given level of accuracy is unaffected by the model structure.

### 2.3. Parameter estimation with the approximated likelihood

To estimate the unknown parameters, the gradient of the approximated log-likelihood is needed. We calculate the gradient $\boldsymbol{\nabla}_{\boldsymbol{\theta}}$ of Equations (7) and (11) to obtain, for each $\theta \in \boldsymbol{\theta}$, and for each Method $\in$ {Lap2, AGHQ},

$$\nabla_\theta = \sum_{g=1}^{G}\sum_{f=1}^{N_g}\left(\frac{\partial l_{fg}^{\text{Method}}(\boldsymbol{\theta}_g|\boldsymbol{y}_{fg})}{\partial\theta} + \frac{\partial\hat{\boldsymbol{z}}_{fg}}{\partial\theta}\frac{\partial l_{fg}^{\text{Method}}(\boldsymbol{\theta}_g|\boldsymbol{y}_{fg})}{\partial\boldsymbol{z}}\Bigg|^{\boldsymbol{z}=\hat{\boldsymbol{z}}_{fg}^{\text{Method}}}\right), \tag{12}$$

where the second term in the expression is needed since the mode is dependent on the parameter vector $\boldsymbol{\theta}$ (Huber et al., 2004; Jin et al., 2018), and where $\hat{\boldsymbol{z}}_{fg}^{\text{Lap2}} = \hat{\boldsymbol{z}}_{fg}$ and $\hat{\boldsymbol{z}}_{fg}^{\text{AGHQ}} = \sqrt{2}\boldsymbol{\Gamma}_{fg}\boldsymbol{q}_{j_1,\ldots,j_p} + \hat{\boldsymbol{z}}_{fg}$. Note that AGHQ requires derivatives up to the third-order and the second-order Laplace requires derivatives up to the fifth-order. We analytically derived the derivatives for the generalized partial credit model and nominal response model and give the expressions of the constituents of Equation (12) for each of these in the appendix. The derivatives for the graded response model can be found in the supplementary material of Jin and Andersson (2020). Note that, with the second-order Laplace approximation, the model structure and symmetry of the derivatives also impact the computation of the gradient and just like for the computation of the entries in Equation (8), we compute only the unique entries in Equation (12) and weight them by their frequency.

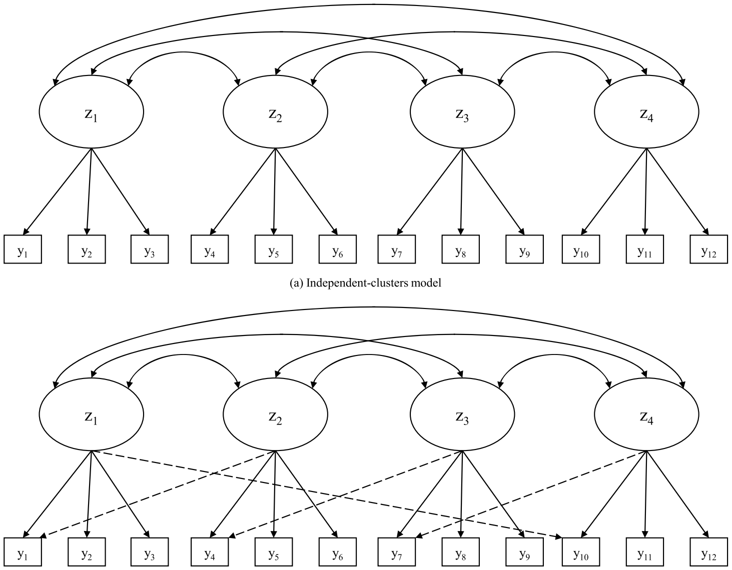
With the gradient, we implement a quasi-Newton method for parameter estimation where the Hessian matrix is approximated with either the empirical cross-product matrix (Berndt et al., 1974) or the Broyden–Fletcher–Goldfarb–Shanno (Nocedal and Wright, 2006, BFGS) method. Let iter denote the iteration number and define $\alpha_{\text{iter}}$ as the step size in the quasi-Newton method. The algorithm proceeds as follows.

1. Let iter $= 0$ and define starting values $\hat{\boldsymbol{\theta}}_{\text{iter}}$.
2. With values $\hat{\boldsymbol{\theta}}_{\text{iter}}$, compute the posterior modes $\hat{\boldsymbol{Z}}$ and the gradient $\boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}_{\text{iter}}}$.
3. Compute the approximated Hessian matrix $\boldsymbol{H}_{\text{iter}}$ from the gradient $\boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}_{\text{iter}}}$.
4. Update the parameter estimates with $\hat{\boldsymbol{\theta}}_{\text{iter}+1} = \hat{\boldsymbol{\theta}}_{\text{iter}} + \alpha_{\text{iter}} \times \boldsymbol{H}_{\text{iter}}^{-1}\boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}_{\text{iter}}}$ and let iter $=$ iter $+ 1$.
5. Repeat steps 2-4 until $\max|\hat{\boldsymbol{\theta}}_{\text{iter}} - \hat{\boldsymbol{\theta}}_{\text{iter}-1}| < \text{TOL}$.

We propose using starting values $a_{ijc} = 1.2$ for the slope parameters, an even sequence from $m_i - 2$ to $-(m_i - 2)$ for the intercept parameters (hence, starting value 0 if $m_i = 2$ and starting values 1 and -1 if $m_i = 3$) and $\sigma_{jk} = 0.5$, with TOL $=$ 0.0001 and $\alpha_{\text{iter}} = 1.0$ as default settings. If $\max|\boldsymbol{H}_{\text{iter}}^{-1}\boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}_{\text{iter}}}| > 0.25$, we suggest to instead set $\alpha_{\text{iter}} = 0.25/\max|\hat{\boldsymbol{\theta}}_{\text{iter}} - \hat{\boldsymbol{\theta}}_{\text{iter}-1}|$ to avoid changing the parameter estimates too much in each iteration. Note that we directly maximize the approximated marginal log-likelihood function instead of using the EM (Dempster et al., 1977) algorithm. As a result, $\hat{\boldsymbol{z}}_{fg}$ is treated as a function of $\theta$ as implied in Equation (12).

### 2.4. Inference with the approximated likelihood

To draw inference we suggest using the inverse of the observed information matrix. The observed information matrix can be approximated by a numerical approximation to the Jacobian of the observed gradient in Equation (12) or the approximation from the BFGS algorithm. The results in Andersson and Xin (2021) indicated that using the numerical approximation to the Jacobian was accurate with correctly specified independent-clusters models and we therefore use this method in the current study. The numerical approximation to the Jacobian is obtained by defining an objective function with the unknown parameters as a vector-valued input argument, which computes and returns the exact observed gradient of the approximated log-likelihood, given in Equation (12). Before computing the gradient, the objective function updates the mode for each response pattern based on the parameters of the input argument. The Jacobian of this function is then approximated with a finite difference approach as implemented in the R package numDeriv (Gilbert and Varadhan, 2019). This method thus provides an approximation of the second derivatives of the approximated log-likelihood, taking the mode estimation into account when doing so. It is also possible to use the sandwich estimator, based on an approximation of the observed information matrix and the empirical cross-product matrix, to obtain robust standard errors.



(a) Independent-clusters model

(b) Cross-loading model

**Fig. 2.** Illustration of the two models with 12 observed variables used in the simulation study.

## 3. Simulation study

We performed a simulation study to investigate the parameter recovery and the computational efficiency of the second-order Laplace approximation (Lap2) method for multiple group models by comparing with the first-order Laplace approximation (Lap1) and adaptive Gauss-Hermite quadrature with three (AGHQ3) and five (AGHQ5) quadrature points. The R package lamle (Andersson and Jin, 2022) was used for parameter estimation with all methods.

### *3.1. Simulation design*

In the simulations, we considered models with four latent variables and four groups, with 500 participants per group. Two settings were manipulated in the simulation design: (1) the type of multidimensional model - an independent-clusters model or a cross-loading model, and (2) the number of observed variables - three or four per dimension, for a total of either 12 or 16 observed variables. We considered only binary observed variables because the Laplace approximation has been shown in previous research to perform the worst in this setting (Joe, 2008; Andersson and Xin, 2021). We considered only the case of a low number of observed variables for the same reason. The conditions of experimental setting (1) represent between-item and within-item multidimensional models, respectively (Wang et al., 2004). In the between-item case, each observed variable is assumed to measure a single latent variable while in the within-item case some observed variables measure more than one latent variable via cross-loadings. Specifically, in the scenario with 12 observed variables, we added cross-loadings to one observed variable for each latent variable, resulting in four cross-loadings. Similarly, two observed variables in each dimension load on another dimension in the scenario with 16 observed variables, resulting in eight cross-loadings. The two models in the setting with 12 observed variables are presented in Fig. 2, where the ellipses represent the latent variables, the rectangles represent the observed variables, the solid lines with two arrows represent covariances between the latent variables, and the solid lines with one arrow represent the main loadings, with cross-loadings represented by dotted lines with one arrow. For both types of models, the covariances for the latent variables are freely estimated in each group and the mean vector and variances for the latent variables are freely estimated in each group except the first one, where the means and variances are all fixed to 0 and 1, respectively. The manipulation of the simulation settings led to $2 \times 2 = 4$ conditions. 1000 replications were conducted under each condition.

Data were generated in R version 4.1.1 (R Core Team, 2021). We simulated binary data using the graded response model with slope and intercept parameters given in Tables A.4-A.6 in the Appendix. We selected these parameters to have a setting which closely resembles real-life examples in educational and psychological measurement (Ayala, 2009). The latent variables were generated from a multivariate normal distribution with the mean vectors $(-1, -1, -1, -1)$, $(-.5, -.5, -.5, -.5)$, $(0, 0, 0, 0)$, and $(.25, .25, .25, .25)$ in the respective group, chosen to represent common differences in proficiency between age groups in practice. The covariances, which were identical for each group, were set to values between 0.4 and 0.6 and

**Table 1**
Convergence rates (in percent), average absolute bias, average coverage rate of 95% confidence intervals (in percent), average root mean squared error, average estimation time (in seconds), and average number of iterations for the four-dimensional multiple group models with sample size 2000 and different numbers of variables.

| Outcome measure | Model | J | Lap1 | Lap2 | AGHQ3 | AGHQ5 |
|---|---|---|---|---|---|---|
| Convergence rate | Independent-clusters model | 12 | 100 | 100 | 100 | 100 |
| | | 16 | 100 | 100 | 100 | 100 |
| | Cross-loading model | 12 | 67.6 | 100 | 98.9 | 99.3 |
| | | 16 | 99.7 | 100 | 100 | 100 |
| Average absolute bias | Independent-clusters model | 12 | 0.076 | 0.013 | 0.024 | 0.011 |
| | | 16 | 0.050 | 0.010 | 0.017 | 0.009 |
| | Cross-loading model | 12 | 0.080 | 0.015 | 0.022 | 0.012 |
| | | 16 | 0.048 | 0.010 | 0.016 | 0.009 |
| Average coverage rate | Independent-clusters model | 12 | 90.0 | 94.6 | 94.6 | 94.8 |
| | | 16 | 92.1 | 95.0 | 94.9 | 95.0 |
| | Cross-loading model | 12 | 91.3 | 94.7 | 94.8 | 94.9 |
| | | 16 | 92.4 | 94.8 | 94.7 | 94.8 |
| Average root mean squared error | Independent-clusters model | 12 | 0.040 | 0.027 | 0.025 | 0.027 |
| | | 16 | 0.025 | 0.020 | 0.019 | 0.020 |
| | Cross-loading model | 12 | 0.047 | 0.030 | 0.029 | 0.031 |
| | | 16 | 0.029 | 0.024 | 0.023 | 0.024 |
| Average estimation time | Independent-clusters model | 12 | 23.71 | 31.64 | 271.15 | 2182.70 |
| | | 16 | 29.32 | 40.17 | 363.10 | 2987.91 |
| | Cross-loading model | 12 | 78.00 | 88.22 | 372.17 | 2437.80 |
| | | 16 | 40.91 | 164.52 | 475.47 | 3365.19 |
| Average number of iterations | Independent-clusters model | 12 | 30.77 | 29.93 | 29.75 | 29.83 |
| | | 16 | 30.67 | 30.32 | 30.20 | 30.31 |
| | Cross-loading model | 12 | 113.63 | 35.82 | 42.14 | 34.14 |
| | | 16 | 38.09 | 31.63 | 31.89 | 31.56 |

*Notes.* J = number of observed variables, Lap1 = first-order Laplace, Lap2 = second-order Laplace, AGHQ3/AGHQ5 = adaptive Gauss-Hermite quadrature with 3 or 5 quadrature points.

are given in Table A.7 in the Appendix. The variances for the latent variables were fixed to one. Hence, the four groups varied in the means of the latent variables but shared the same correlations among the four dimensions. Note that we freely estimated the covariances in each group.

With the generated data, we employed the Lap1, Lap2, AGHQ3, and AGHQ5 estimation methods. To assess the performance of the four methods, we examined their statistical properties in terms of convergence rate, parameter recovery, and computational speed. Successful convergence was determined by fulfilling all of the following three criteria: 1) the algorithm stopped within 500 iterations, 2) the empirical cross-product matrix was positive definite, and 3) the approximated observed information matrix was positive definite. After concluding the simulation, we also inspected the parameter estimates and standard errors with each method to detect outlying replications. We computed the convergence rate in percent for each method and setting. Regarding parameter recovery, for a parameter $\theta$ with the estimate $\hat{\theta}^r$ in replication $r$, define the absolute bias as $|\text{bias}|_\theta = |\sum_{r=1}^R(\hat{\theta}^r - \theta)/R|$ and the root mean squared error (RMSE) as $\text{RMSE}_\theta = \sqrt{\sum_{r=1}^R(\hat{\theta}^r - \theta)^2/R}$. We computed overall measures of the recovery of the parameters (slopes, intercepts, covariances, variances, and means) in terms of average absolute bias, average RMSE, and average coverage rate of 95% confidence intervals estimated with the standard errors from the observed information matrix, across all parameters in a given setting. To evaluate the computational efficiency of the four estimation methods, we recorded the time information and the number of iterations required in the estimation. The computational efficiency is comparable between the methods since all methods are based on the same code base written in C++. The total simulation time exceeded 3600 core hours.

### *3.2. Results*

In this subsection, we present the results of the simulation study. First in Table 1 are the convergence rates of Lap1, Lap2, AGHQ3, and AGHQ5 estimation methods for the four-dimensional multiple group models. It suggests that all the estimation methods attained 100% convergence for the independent-clusters models. However, with cross-loading models, the Lap2 method outperformed the other methods and was the only method to reach a 100% convergence rate. The Lap1 method was particularly problematic with respect to convergence (convergence rate = 67.6%) in the cross-loading scenario with 12 observed variables. We excluded the non-converged replications in subsequent comparisons of the methods.

Next, we evaluate the recovery of parameters and summarize the results of Table 1 concerning this. With respect to average absolute bias, Lap2 and AGHQ5 estimation methods produced less bias compared with Lap1 and AGHQ3. With an

**Table 2**
Number of unique nonzero 3rd and 4th derivatives and the number of unique nonzero entries in the sums of the second-order Laplace approximation for the four different models considered in the simulation.

| Model | J | Unique 3rd | Unique 4th | Unique sum A | Unique sum B | Unique sum C |
|---|---|---|---|---|---|---|
| Independent-<br>clusters model | 12<br>16 | 4<br>4 | 4<br>4 | 4<br>4 | 10<br>10 | 10<br>10 |
| Cross-loading<br>model | 12<br>16 | 12<br>16 | 16<br>22 | 20<br>28 | 170<br>322 | 114<br>214 |

*Note.* J = number of observed variables.

increasing number of observed variables, the average absolute bias decreased, especially for the Lap1 estimation method. The types of model, independent-clusters or cross-loading, did not impact estimation accuracy much. Regarding average coverage of 95% confidence intervals, Lap2, AGHQ3, and AGHQ5 showed similar results and performed better than Lap1. In addition, the manipulation settings barely had influence on this evaluation criterion. Besides the accuracy of parameter estimates, we also considered the estimation precision. The results suggest that compared with Lap1, other estimation methods produced more precise estimates, especially when the number of observed variables was 12. Increasing the number of observed variables or using a simpler model improved the average RMSE for all estimation methods. In general, the Lap2 estimation method exhibited similar estimation accuracy and precision with the AGHQ5 method, outperforming the AGHQ3 method and especially the Lap1 method.

Regarding computational efficiency, overall, the Lap1 method cost the least time (23.71 to 78.00 seconds per replication on average), followed by the Lap2 method (31.64 to 164.52 seconds), while the adaptive Gauss-Hermite quadrature methods required much longer time - over 270 seconds for AGHQ3 and more than 2100 seconds for AGHQ5 for all settings. Regarding the influence of the experimental factors, all the methods needed more time under the cross-loading conditions compared with the independent-clusters conditions. The computational time increased as the number of observed variables rose, except for the Lap1 method under the cross-loading conditions. The reason for this decrease is the need for additional iterations in the algorithm when using the Lap1 method in the cross-loading setting with 12 observed variables compared to 16 observed variables. It is worth mentioning that the simulation study in Bianconcini (2014) showed that Lap2 needs substantially more steps than AGHQ5, whereas our results show that they tend to converge within similar number of iterations.

To explain these computational results, it is useful to consider their relationship to the expression of the second-order Laplace approximation in terms of the unique quantities that are needed and the size of the sums included in the approximation. With an arbitrary four-dimensional model, the number of unique third derivatives of $h_{fg}(\boldsymbol{z})$ is at most 20 and the number of unique fourth derivatives of $h_{fg}(\boldsymbol{z})$ is at most 35. The sums in Equation (8) consist of entries that are products of these derivatives and the entries of the inverse of $\boldsymbol{H}_{fg}$. When ignoring symmetries and zero entries, sum A with fourth-order derivatives has $4^4 = 256$ entries, and sums B and C with third-order derivatives each have $4^6 = 4096$ entries.

In Table 2, we present the number of unique derivatives that must be computed for each model we considered and the number of unique entries required in the sums of Equation (8). The number of unique entries does not change when increasing the number of observed variables for the independent-clusters model because the model structure remains the same. This means that the computational time is essentially a linear function of the number of observed variables. However, since the model structure changes for the cross-loading model when increasing the number of observed variables from 12 to 16, there is an increased number of unique entries both for the derivatives and for the resulting sums. Nevertheless, the reduction in the number of entries is substantial compared to the unfiltered number since at most 28 out of the total 256 of sum A, 322 of the total 4096 of sum B, and 214 of the total 4096 of sum C have to be computed. This illustrates that the number of derivatives that must be computed for these models is still quite small compared the total which explains the high computational efficiency of the second-order Laplace approximation. Note that the function $h_{fg}$ must be evaluated only once for each unique response pattern in the data for either the first- or second-order Laplace approximation whereas with adaptive quadrature it needs to be evaluated either 81 (AGHQ3) or 625 (AGHQ5) times for each unique response pattern in the data. As seen in the simulation study, this results in drastically increased computational time for adaptive quadrature relative to the Laplace approximations for the four-dimensional models considered.

In sum, the simulation study suggests that the Lap2 estimation method led to desirable outcomes in three aspects. First, unlike the Lap1 method, which suffered from non-convergence problems under some conditions, the Lap2 method achieved convergence in all the simulated data sets. Second, the Lap2 method yielded both accurate and precise parameter estimates, which was comparable to AGHQ5. Both Lap2 and AGHQ5 methods outperformed Lap1 and AGHQ3 methods in terms of parameter recovery. Third, the Laplace approximation methods greatly improved the computational speed compared to the adaptive Gauss-Hermite quadrature methods. Thus, we conclude that the Lap2 estimation method can produce satisfactory parameter estimates with a substantial improvement of computational efficiency compared to adaptive Gauss-Hermite quadrature for estimation of multidimensional multiple group models.

**Table 3**
Estimated latent means (se) of 2009 PISA mathematics literacy, reading literacy, and science literacy in Hong Kong (the reference group), Macao, Shanghai, and Chinese Taipei.

| | Mathematics | Reading | Science |
|---|---|---|---|
| Hong Kong | 0 | 0 | 0 |
| Macao | -0.31 (0.03) | -0.58 (0.02) | -0.45 (0.02) |
| Shanghai | 0.47 (0.03) | 0.20 (0.03) | 0.25 (0.03) |
| Chinese Taipei | -0.07 (0.03) | -0.50 (0.02) | -0.32 (0.03) |

## 4. Empirical illustration

### *4.1. Data and models*

To illustrate the proposed estimation method and compare against alternatives, we utilized data from Hong Kong, Macao, Shanghai and Chinese Taipei in the 2009 Programme for International Student Assesment (Schleicher et al., 2009, PISA). PISA is a large-scale educational assessment run by the Organisation for Economic Co-operation and Development (OECD) which aims to measure student achievement in mathematics, reading, and science, and monitor the outcomes of education systems internationally. We estimated three-dimensional multiple group (i.e., four regions) independent-clusters graded response models, where the dimensions corresponded to mathematics literacy, reading literacy and science literacy. The total number of respondents were 21690 but we removed 18 of these due to excessive numbers of missing values. The sample sizes in each region were 4792 in Hong Kong, 5948 in Macao, 5113 in Shanghai and 5819 in Chinese Taipei. There were 35 mathematics items, 100 reading items and 53 science items for a total of 188 items, out of which 176 items were binary scored and 12 were scored in three categories. Due to the PISA 2009 sampling design, each respondent only answered a subset of the total items included in the study and missing values are assumed missing at random. The average number of responses to the 188 total items was 54.55 across all regions. We simultaneously estimated the item parameters and the mean vectors and covariance matrices in each region using the item response data. Hong Kong was set as the reference group and the means and variances of the latent variable were not estimated in this group. The item parameters were considered invariant between regions. The model had nine free mean parameters, nine free variance parameters, 12 free covariance parameters, and 388 free item parameters, for a total of 418 parameters that were uniquely estimated.

### *4.2. Estimation settings*

Estimation was done by maximizing the approximated log-likelihood, where the first-order Laplace (Lap1), second-order Laplace (Lap2) and adaptive Gauss-Hermite quadrature with 3, 5 and 13 quadrature points (AGHQ3, AGHQ5, and AGHQ13) were used for the integral approximations. The most accurate method out of these according to underlying theory is the AGHQ13 method, which has a fifth-order accuracy (Jin and Andersson, 2020). We thus used this as the reference method to compare the other methods against. The methods used the same starting values and estimation settings: tolerance of 0.0001, step size 0.5 for the first 25 iterations, maximum update direction 0.25, and maximum number of iterations 500. The R package lamle was used for all methods. All methods converged successfully after 36 iterations.

### *4.3. Results*

#### *4.3.1. Estimated distribution parameters*

The estimated mean vectors and associated standard errors from the Lap2 method are shown in Table 3, where Hong Kong is set as the reference group and thus has a mean vector of 0 and variances equal to 1. The estimated means indicate that Shanghai is the highest performing region overall. The performance in reading and science are the highest in Hong Kong and Shanghai. Macao and Chinese Taipei have similar profiles, with slightly lower mean estimates compared to Hong Kong in all domains. The estimated covariance matrix for the latent mathematics, reading, and science literacy indicate that the domains are highly correlated, with estimated correlations between 0.86 and 0.92 in the reference group Hong Kong. The results for the three other regions were similar to Hong Kong and the full results are provided as covariance matrices in the Appendix.

#### *4.3.2. Computational efficiency and accuracy*

The log-likelihood values, estimation times and the parameter estimates were obtained after convergence. The log-likelihood values and estimation times are given in Table 4. These statistics reflect the results from the simulation study, in that the Lap1 method is the fastest but which approximates the log-likelihood the worst. Meanwhile, the Lap2 method is almost as fast as Lap1 while having a log-likelihood value that is very close to AGHQ13. Compared to AGHQ3, Lap2 improves both on the accuracy in terms of the log-likelihood value and the computational efficiency. AGHQ5 gives a log-likelihood value that is the closest to the reference AGHQ13, but the difference to Lap2 is small. The speed improvement of the Laplace-based methods relative to the adaptive quadrature methods is not as great for this three-dimensional model because of the

**Table 4**
Log-likelihood values and estimation times in seconds for five estimation methods with the 2009 PISA data from Hong Kong, Macao, Shanghai, and Chinese Taipei.

| | Lap1 | Lap2 | AGHQ3 | AGHQ5 | AGHQ13 |
|---|---|---|---|---|---|
| Log-likelihood | -627557.7 | -627402.9 | -627426.6 | -627405.3 | -627404.7 |
| Estimation time | 532.3 | 608.5 | 1800.7 | 6031.8 | 99081.1 |

*Note.* Lap1 = first-order Laplace, Lap2 = second-order Laplace, AGHQ3/AGHQ5/AGHQ13 = adaptive Gauss-Hermite quadrature with 3, 5, or 13 quadrature points.

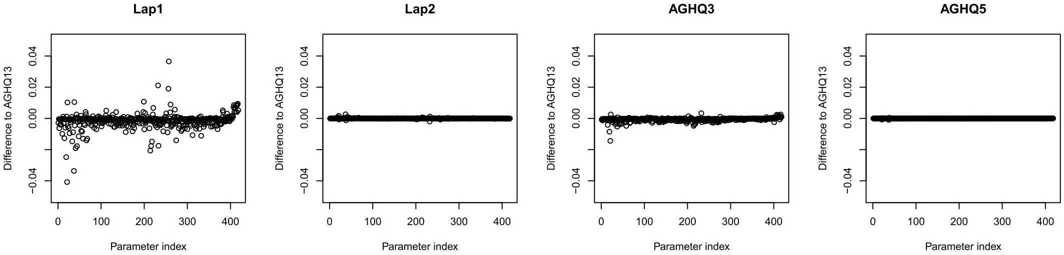


**Fig. 3.** Differences between parameter estimates obtained from Lap1, Lap2, AGHQ3, and AGHQ5, compared to parameter estimates from AGHQ13.

fewer number of total quadrature points needed with this model compared to a four-dimensional model. Nevertheless, the Lap2 method was almost three times faster than AGHQ3 and almost 10 times faster than AGHQ5 in estimation.

The log-likelihood values suggested that the methods provided slightly different results but this does not necessarily indicate if there are substantial differences in the parameter estimates from the different methods. To illustrate potential differences between the methods, we plotted the differences in the parameter estimates for Lap1, Lap2, AGHQ3 and AGHQ5 when compared to the AGHQ13 method. These differences are provided in Fig. 3, showing that Lap1 has the largest differences to the AGHQ13 method, followed by AGHQ3, Lap2, and AGHQ5. Overall, the differences in parameter estimates are small for all methods, differing at most by 0.0408 in absolute value for Lap1, 0.0027 in absolute value for Lap2, 0.0144 in absolute value for AGHQ3, and 0.0009 in absolute value for AGHQ5. This indicates that all of the methods can be considered sufficiently accurate in terms of parameter recovery in this particular setting, which is a consequence of the fairly large number of item responses by each student.

## 5. Discussion

Estimation of generalized linear latent variable models is computationally demanding in high dimensions which hinders their usage in many practical situations. In this study we implemented a second-order Laplace approximation method for estimation of generalized linear latent variable models with categorical observed variables and multiple groups. The practical consequences of our results are that the application of generalized linear latent variable models with high dimensionality is possible to do in situations with large sample sizes and with many parameters and that estimation time is reduced compared to alternative methods in other settings. In the numerical illustration, the second-order Laplace approximation method was highly computationally efficient compared to adaptive quadrature with three and five quadrature points per dimension. Meanwhile, the estimation accuracy was improved in relation to the first-order Laplace approximation and adaptive quadrature with three quadrature points while estimation accuracy was almost identical to that attained with five quadrature points. Since the second-order Laplace approximation has the same theoretical error rate as adaptive Gauss-Hermite quadrature with four to six quadrature points (Jin and Andersson, 2020), the second-order Laplace approximation was substantially more efficient than adaptive Gauss-Hermite quadrature at the same level of theoretical accuracy with the examples used in this study.

The results of this study thus imply that the Laplace approximation has a computational efficiency far above that of adaptive quadrature using a number of quadrature points which provides the same theoretical error rate. For the four-dimensional models considered here, the second-order Laplace approximation was also more efficient than adaptive quadrature with three quadrature points, which has a lower theoretical error rate than the second-order Laplace approximation. Generally speaking, the efficiency advantage will be lower with fewer latent variables and the efficiency advantage will be greater with more latent variables. The computational advantage of the second-order Laplace approximation is however also dependent on the complexity of the model structure. The highest efficiency gains are realized when each observed variable is related to only a single latent variable out of many correlated latent variables. The lowest gains are realized when specifying an unrestricted model with uncorrelated latent variables.

Our results can guide the practical use of latent variable models in the following ways. First, compared to the regular Laplace approximation the second-order Laplace approximation is preferred for most situations due to the added estimation accuracy. An exception is for the case of complex models with many indicators where the second-order method may still be too time-consuming to utilize. With the most commonly used model structures, the second-order Laplace approximation is however fast enough to support up to 12 correlated latent variables (Andersson and Xin, 2021) which should cover most settings in practice. Second, we argue that the second-order Laplace approximation is especially useful for cases when adaptive quadrature with four or more quadrature points is impossible to practically conduct. If adaptive quadrature with many quadrature points is possible to employ, the second-order Laplace approximation becomes less suitable since a higher accuracy can be attained with adaptive quadrature by increasing the number of quadrature points.

Compared to alternatives such as adaptive quadrature and simulation-based methods, the second-order Laplace is attractive because of its high computational efficiency while maintaining a high accuracy. It is also attractive relative to the simulation-based methods due to the highly efficient computations of the log-likelihood and the observed information matrix, which can be extremely time-consuming for simulation-based methods when the sample size is large. Downsides to using the second-order Laplace approximation are that the computational advantage reduces for complex models and that the method requires the computation of a considerable amount of higher-order derivatives that do not easily generalize for different measurement models. It is also not straightforward with the Laplace approximation or adaptive quadrature to utilize the independence structure of the latent variables to improve the efficiency, as is possible with regular numerical quadrature (Gibbons and Hedeker, 1992; Cai, 2010b). However, regular numerical quadrature is unfeasible when the number of correlated latent variables is larger than three.

Future avenues of research include supporting additional types of observed variables combined with the categorical observed variables considered here. For example, providing an efficient yet accurate method that supports combinations of continuous data, count data, ordinal data and nominal data would be ideal to have. In addition, extensions of the approach to support mixture models and additional latent variable distributions beyond the normal distribution are possible.

## Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2023.107710.

## References

Andersson, B., Jin, S., 2022. lamle: maximum likelihood estimation of latent variable models using adaptive quadrature and Laplace approximations. https://github.com/bjoernhandersson/lamlepub/releases/tag/v0.1.2-alpha.

Andersson, B., Xin, T., 2021. Estimation of latent regression item response theory models using a second-order Laplace approximation. J. Educ. Behav. Stat. 46 (2), 244–265. https://doi.org/10.3102/1076998620945199.

Ayala, R., 2009. The Theory and Practice of Item Response Theory, Methodology in the Social Sciences. The Guilford Press, New York, NY.

Berndt, E.R., Hall, B.H., Hall, R.E., Hausman, J.A., 1974. Estimation and inference in nonlinear structural models. Ann. Econ. Soc. Meas. 3 (4), 653–665.

Bianconcini, S., 2014. Asymptotic properties of adaptive maximum likelihood estimators in latent variable models. Bernoulli 20 (3), 1507–1531. https://doi.org/10.3150/13-BEJ531.

Bianconcini, S., Cagnone, S., 2012. Estimation of generalized linear latent variable models via fully exponential Laplace approximation. J. Multivar. Anal. 112, 183–193. https://doi.org/10.1016/j.jmva.2012.06.005.

Bock, R.D., 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37 (1), 29–51. https://doi.org/10.1007/BF02291411.

Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 46 (4), 443–459. https://doi.org/10.1007/BF02293801.

Cagnone, S., Monari, P., 2013. Latent variable models for ordinal data by using the adaptive quadrature approximation. Comput. Stat. 28 (2), 597–619. https://doi.org/10.1007/s00180-012-0319-z.

Cai, L., 2010a. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. J. Educ. Behav. Stat. 35 (3), 307–335. https://doi.org/10.3102/1076998609353115.

Cai, L., 2010b. A two-tier full-information item factor analysis model with applications. Psychometrika 75 (4), 581–612. https://doi.org/10.1007/s11336-010-9178-0.

Cho, A.E., Wang, C., Zhang, X., Xu, G., 2021. Gaussian variational estimation for multidimensional item response theory. Br. J. Math. Stat. Psychol. 74 (S1), 52–85. https://doi.org/10.1111/bmsp.12219.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., Ser. B, Methodol. 39 (1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Gibbons, R.D., Hedeker, D.R., 1992. Full-information item bi-factor analysis. Psychometrika 57 (3), 423–436. https://doi.org/10.1007/BF02295430.

Gilbert, P., Varadhan, R., 2019. numDeriv: accurate numerical derivatives, r package version 2016.8-1.1. https://CRAN.R-project.org/package=numDeriv.

Huber, P., Ronchetti, E., Victoria-Feser, M.-P., 2004. Estimation of generalized linear latent variable models. J. R. Stat. Soc., Ser. B, Stat. Methodol. 66 (4), 893–908. https://doi.org/10.1111/j.1467-9868.2004.05627.x.

Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V., Taskinen, S., 2017. Variational approximations for generalized linear latent variable models. J. Comput. Graph. Stat. 26 (1), 35–43. https://doi.org/10.1080/10618600.2016.1164708.

Jin, S., Andersson, B., 2020. A note on the accuracy of adaptive Gauss–Hermite quadrature. Biometrika 107 (3), 737–744. https://doi.org/10.1093/biomet/asz080.

Jin, S., Noh, M., Lee, Y., 2018. H-likelihood approach to factor analysis for ordinal data. Struct. Equ. Model. 25 (4), 530–540. https://doi.org/10.1080/10705511.2017.1403287.

Jin, S., Vegelius, J., Yang-Wallentin, F., 2020. A marginal maximum likelihood approach for extended quadratic structural equation modeling with ordinal data. Struct. Equ. Model. 27 (6), 864–873. https://doi.org/10.1080/10705511.2020.1712552.

Joe, H., 2008. Accuracy of Laplace approximation for discrete response mixed models. Comput. Stat. Data Anal. 52 (12), 5066–5074. https://doi.org/10.1016/j.csda.2008.05.002.

Muraki, E., 1992. A generalized partial credit model: application of an EM algorithm. Appl. Psychol. Meas. 16 (2), 159–176. https://doi.org/10.1177/014662169201600206.

Muthen, B., Lehman, J., 1985. Multiple group IRT modeling: applications to item bias analysis. J. Educ. Behav. Stat. 10 (2), 133–142. https://doi.org/10.3102/10769986010002133.

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., Warton, D.I., 2019. Efficient estimation of generalized linear latent variable models. PLoS ONE 14 (5), 1–20. https://doi.org/10.1371/journal.pone.0216129.

Nocedal, J., Wright, S., 2006. Numerical Optimization. Springer-Verlag, New York, NY.

Noh, M., Lee, Y., 2007. REML estimation for binary data in GLMMs. J. Multivar. Anal. 98 (5), 896–915. https://doi.org/10.1016/j.jmva.2006.11.009.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2004. Generalized multilevel structural equation modeling. Psychometrika 69 (2), 167–190. https://doi.org/10.1007/BF02295939.

Raudenbush, S.W., Yang, M.-L., Yosef, M., 2000. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. J. Comput. Graph. Stat. 9 (1), 141–157. https://doi.org/10.2307/1390617.

Samejima, F., 1969. Estimation of latent ability using a response pattern of graded scores. Psychometrika 34 (Suppl 1), 1–97. https://doi.org/10.1007/BF03372160.

Schilling, S., Bock, R.D., 2005. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. Psychometrika 70 (3), 533–555. https://doi.org/10.1007/s11336-003-1141-x.

Schleicher, A., Zimmer, K., Evans, J., Clements, N., 2009. Pisa 2009 assessment framework: key competencies in reading, mathematics and science. OECD Publishing (NJ1).

Shun, Z., 1997. Another look at the salamander mating data: a modified Laplace approximation approach. J. Am. Stat. Assoc. 92 (437), 341–349. https://doi.org/10.1080/01621459.1997.10473632.

Thomas, N., 1993. Asymptotic corrections for multivariate posterior moments with factored likelihood functions. J. Comput. Graph. Stat. 2 (3), 309–322.

Wang, W.-C., Chen, P.-H., Cheng, Y.-Y., 2004. Improving measurement precision of test batteries using multidimensional item response models. Psychol. Methods 9 (1), 116–136. https://doi.org/10.1037/1082-989X.9.1.116.

Zhu, J., Eickhoff, J., Yan, P., 2005. Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. Biometrics 61 (3), 674–683. https://doi.org/10.1111/j.1541-0420.2005.00343.x.

Paper IV

# Estimation of generalized linear latent variable models for performance and process data with ordinal, continuous, and count observed variables

**Maoxin Zhang, Björn Andersson, Shaobo Jin**

Submitted to *British Journal of Mathematical and Statistical Psychology*.

# Estimation of generalized linear latent variable models for performance and process data with ordinal, continuous, and count observed variables

Maoxin Zhang[1], Björn Andersson[1], and Shaobo Jin[2]

[1]Center for Educational Measurement, University of Oslo

[2]Department of Mathematics, Uppsala University

## Author Note

Correspondence: Björn Andersson https://orcid.org/0000-0002-9007-2440, bjorn.andersson@cemo.uio.no. All authors declare that they have no conflicts of interest.

## Abstract

A collection of different data types often occurs in psychological and educational measurement such as computer-based assessments that record performance and process data (e.g., response times and the number of actions). Modeling such data requires specific models for each data type and accommodating complex dependencies between multiple variables. Generalized linear latent variable models are suitable for modeling mixed data simultaneously, but estimation can be computationally demanding. A fast solution is to use Laplace approximations but existing implementations of joint modeling of mixed data types are limited to ordinal and continuous data. To address this limitation, we derive an efficient estimation method to simultaneously model ordinal data, continuous data, and count data using first-order and second-order Laplace approximations. We illustrate the approach with an empirical example and conduct simulations to evaluate the performance of the proposed method in terms of estimation efficiency, convergence, and parameter recovery. The results suggest that the second-order Laplace approximation achieves a higher convergence rate and produces accurate yet fast parameter estimates compared to the first-order Laplace approximation, while the time cost increases with the complexity of model specification. Additionally, models that consider the dependence of the variables from the same stimulus better fit the empirical data.

*Keywords:* mixed data types; Laplace approximation; estimation efficiency; high dimensionality

# Estimation of generalized linear latent variable models for performance and process data with ordinal, continuous, and count observed variables

## Introduction

Due to technological advances in data collection via digital devices, a mixture of continuous and discrete data (e.g., binary, categorical, and count) frequently occurs in assessment contexts (De Leon & Chough, 2013, Chapter 1). Compared to traditional paper-and-pencil tests that only record the final answer to the items, computer-based assessments can track the entire human-computer interaction sequence (e.g., mouse clicks and keyboard input with timestamps) and compile such information in log files. From log files, researchers can extract different types of indicators for further analyses, such as scored responses (categorical), response time spent on single items (continuous), response time until the first action (continuous), and the number of actions for each item (count). Such information is widely available in large-scale assessments, providing abundant research materials for understanding participants' task-taking behaviors (De Boeck & Scalise, 2019; Ulitzsch, von Davier, & Pohl, 2020). These data also commonly exist in game-based assessments (Landers, Armstrong, Collmus, Mujcic, & Blaik, 2021), which routinely collect the number of correct or incorrect trials, the number of mouse clicks, and the performance scores. In addition, sophisticated measurement tools such as eye-tracking devices also produce mixed data such as fixation count and fixation duration (Steinfeld, 2016). Hence, a combination of continuous and discrete data widely exists in educational and psychological assessments, providing researchers and practitioners with valuable information on diverse aspects of the respondents.

However, a mixture of different types of data poses challenges for conventional statistical methods because of the complex dependence structures that often exist (De Leon & Chough, 2013, Chapter 1). To be specific, dependence can stem from a) the same type of indicator like the responses to a number of items, and b) different types of indicators based on the same stimulus such as the item response and the response time from the same task. The former type of dependence is typically handled by introducing latent variables, while the latter type is often ignored. However, ignoring the dependence of indicators from the same task can lead to biased parameter estimation (De Boeck & Scalise, 2019; Meng, Tao, & Chang, 2015). Additionally, the inherent non-normality of categorical and count data means that traditional analysis methods that assume continuous and normally distributed observed variables are less suitable to use.

Despite the above-mentioned challenges, multiple approaches to handling the issue of a mixture of different types of data exist. Among them, drawing inferences for each type of measure via separate models is the simplest approach. For example, researchers can analyze ordinal performance data via item response theory models, continuous response time via factor analysis, and the number of actions via count data models. However, a multiple testing issue arises (De Leon & Chough, 2013, Chapter 2) and the approach cannot capture relationships between the measures because they are modeled separately. Therefore, a single multivariate model is regarded as a more appealing approach. To estimate such models with traditional methods, data may be converted into the same type by recoding continuous data as categorical data according to certain cutoff values or treating discrete data as continuous. The former method causes a loss of information while the latter violates a model assumption. Neither of these approaches are ideal and it is instead recommended to treat the observed variables as they are (Huber, Ronchetti, & Victoria-Feser, 2004). For joint analysis of ordinal and continuous data, limited-information estimation with polyserial correlations may be used (Olsson, Drasgow, & Dorans, 1982). However, such a method cannot handle count data and the existence of missing data poses an issue in estimation.

An alternative approach is therefore to model mixed data jointly under the framework of generalized linear latent variable models (GLLVMs; Bartholomew, Knott, & Moustaki, 2011; Huber et al., 2004; Rabe-Hesketh, Skrondal, & Pickles, 2004). A complicating factor for GLLVMs concerns the estimation of model parameters. Typically, full-information maximum likelihood or Bayesian estimation has been proposed. Bayesian inference is based on the posterior distribution of the freely estimated parameters given the data and priors of the parameters (Bartholomew et al., 2011, *p.* 30). When the dimension is high or models are very complex, Markov Chain Monte Carlo (MCMC) methods are often used. For example, Man and Harring (2022) and Qiao, Jiao, and He (2022) jointly modeled ordinal, continuous, and count data with Bayesian methods. MCMC methods are computationally demanding with many latent variables and residual dependence between multiple observed variables related to the same stimulus or task has therefore commonly been ignored when using Bayesian methods (Man & Harring, 2022; Qiao et al., 2022; Ulitzsch et al., 2020).

In contrast to Bayesian estimation, full-information maximum likelihood integrates out the latent variables from the likelihood function and maximizes the marginal likelihood. However, the integrals do not have closed-form solutions for GLLVMs and approximation methods are required to compute them. One approach is Gauss-Hermite quadrature (GHQ) which has been implemented for GLLVMs with a collection of data from different distributions in the exponential family (Moustaki, 1996; Moustaki & Knott, 2000). GHQ works well for simple models but becomes unfeasible with more than three latent variables because the computational cost grows exponentially as the latent variable dimension increases (Andersson & Xin, 2021; Huber et al., 2004). Adaptive Gauss-Hermite quadrature (AGHQ) identifies integration intervals with rapid changes and reduces the required number of quadrature points (Rabe-Hesketh, Skrondal, & Pickles, 2002). AGHQ methods for generalized linear latent and mixed models are available in a Stata package *gllamm* (Rabe-Hesketh et al., 2004) and both quadrature methods are available in M*plus* (Muthén & Muthén, 2017). Although AGHQ is faster than GHQ, it is still computationally demanding when the dimension is high. Instead, methods using Laplace approximations are promising to approximate the required integrals accurately and fast (Andersson & Xin, 2021; Huber et al., 2004; Niku, Warton, Hui, & Taskinen, 2017).

First-order Laplace approximations have been proposed to estimate GLLVMs for mixed data with distributions in the exponential family (Huber et al., 2004). Estimation with first-order Laplace approximations has been implemented in the R package *gllvm* (Niku et al., 2017) but supports only one type of indicator at a time. It is worth noting that first-order Laplace approximations (Lap1) are equivalent to AGHQ with one quadrature point per dimension when using the posterior mode and Hessian, and the method thus works highly efficiently with complex, high-dimensional models. However, this comes at the cost of non-convergence and inaccuracy with binary data and few observed variables (Andersson & Xin, 2021; Joe, 2008). To handle issues regarding convergence and accuracy in parameter recovery, a second-order Laplace approximation (Lap2) can be used (Shun, 1997). Lap2 requires higher-order derivatives to obtain a more accurate approximation by including more information but with the downside that it needs more time in estimation (Andersson, Jin, & Zhang, 2023). A recent R package called *lamle* (Andersson & Jin, 2022) has implemented AGHQ, first-order Laplace approximations, and second-order Laplace approximations for categorical data but not for other types of data.

In this article, we propose to apply both first- and second-order Laplace approximations to generalized linear latent variable models with mixed observed variables. Our main interest is to apply Laplace approximations (both Lap1 and Lap2) to enable joint modeling of ordinal, continuous, and count variables based on process data and performance data in educational and psychological measurement, where the residual dependence between observed variables related to the same stimulus is accounted for. This article has three main contributions beyond the existing literature. First, we implement estimation of joint models for count data, continuous data, and ordinal data using Laplace approximations, extending the papers by Huber et al. (2004) and Niku et al. (2017). Second, compared to Huber et al. (2004) and Niku et al. (2017) which only implemented the first-order Laplace approximation, we further implement a second-order Laplace approximation. Third, we provide a comparison between Lap1 and Lap2 in estimating

GLLVMs with a mixture of different observed variables, extending the comparison in Andersson and Xin (2021) and Andersson et al. (2023) from only categorical data to also include continuous and count data.

The remainder of the article is organized as follows. We introduce GLLVMs and derive the estimation algorithm in the Methods section. A motivating example from Programme for International Student Assessment (PISA) is then described to guide our simulation design. Subsequently, simulations are conducted to evaluate estimation of joint models for ordinal, continuous, and count data without or with considering the dependence of indicators from the same task. Finally, the article concludes with a discussion.

## Methods

### Generalized linear latent variable models

GLLVMs are extensions of generalized linear models (Nelder & Wedderburn, 1972, GLMs), which are a class of regression models for discrete or continuous outcomes. GLMs consist of three components (Nelder & Wedderburn, 1972): a) a linear combination of predictors

$$\nu = b + \boldsymbol{\beta}'\boldsymbol{w}, \tag{1}$$

where $b$ and $\boldsymbol{\beta}$ are the intercept and regression coefficients, and $\boldsymbol{w}$ represents $D$-dimensional predictors; b) the outcome variable belonging to an exponential dispersion family; and c) a monotone and differentiable link function $g$ such as the identity, logit, or probit function, which relates the expected value of the outcome variable to the linear combination of predictors $\nu$. In GLMs, there is only one outcome variable and all the variables are observable. When there are multiple correlated indicators that are developed to measure the same construct, such as responses from several cognitive tasks, we can incorporate latent variables to account for the dependence between the indicators. In social science, it is common to develop a battery of tests to measure theoretical constructs since they can not be directly observed.

GLM is extended to generalized linear latent variable models (Bartholomew et al., 2011) by introducing latent variables. Let $y_{if}$ denote the $i$-th observed outcome variable for individual $f$. Following Rabe-Hesketh et al. (2004), a general formula for GLLVMs can be written as

$$g_i(E[y_{if}|\boldsymbol{w}, \boldsymbol{z}]) = b_i + \boldsymbol{\beta}_i^{'}\boldsymbol{w} + \boldsymbol{a}_i^{'}\boldsymbol{z}, \tag{2}$$

where $\boldsymbol{a}_i$ is a vector of slope parameters or factor loadings of item $i$, $\boldsymbol{\beta}_i$ is a $D$-dimensional vector of regression coefficients, and $\boldsymbol{z}$ is a $P$-dimensional vector of latent variables. To link the linear combination and the expected value of the observed variables, the link function $g_i$ must be defined for each observed variable. For the distribution of the latent variables, we assume a multivariate normal distribution. For identification purposes, the means and variances of the latent variables are constrained to zeros and ones, respectively. The observed outcome variables are assumed to be independent conditional on the latent variables (Huber et al., 2004).

Let $\boldsymbol{y}$ be the $I \times 1$-vector of observed outcome variables. The marginal log-likelihood for a response vector $\boldsymbol{y}$ is then

$$l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w}) = \log \int \prod_{i=1}^{I} P_i(\boldsymbol{y}_i|\boldsymbol{w}, \boldsymbol{z})\psi(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{z}, \tag{3}$$

where $\boldsymbol{\theta}$ represents the unknown parameters, $P_i$ defines the measurement model for variable $i$ and $\psi(\cdot)$ is the multivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The latent variables $\boldsymbol{z}$ are unknown and thus need to be integrated out, which requires approximation methods.

## Measurement models

Equation 3 provides a general form of the marginal log-likelihood function for generalized linear latent variable models. As mentioned above, the measurement models $P_i$ and link functions $g_i$ need to be defined according to the specified distribution of the observed variable. Recall that $\boldsymbol{z}$ is the $P$-dimensional vector of latent variables and let $b_i$

be the intercept parameter, $\boldsymbol{a}_i$ be a vector of slope parameters, and $\phi_i$ be the scale parameter, all for the observed variable $i$. Three types of observed data, namely ordinal, continuous, and count data, are considered in this paper and the associated measurement models are given below.

1. *Ordinal responses.* We model the probability of observing each category $c \in 1, \ldots, m_i$ given the latent variables as follows (Muraki, 1992),

$$P_i(y_{if} = c|\boldsymbol{z}, \boldsymbol{w}) = \frac{\exp\left[\sum_{v=1}^{c}(\boldsymbol{a}_i'\boldsymbol{z} + b_{iv} + \boldsymbol{\beta}_i'\boldsymbol{w})\right]}{\sum_{c'=1}^{m_i}\exp\left[\sum_{v=1}^{c'}(\boldsymbol{a}_i'\boldsymbol{z} + b_{iv} + \boldsymbol{\beta}_i'\boldsymbol{w})\right]}, \tag{4}$$

where $b_{iv}$ represents threshold parameters for item $i$ and where a logit link function is assumed.

2. *Continuous responses.* Here we define $P_i(y_{if}|\boldsymbol{z}, \boldsymbol{w})$ as a conditional density function. Following Huber et al. (2004), we assume a normal distribution with an identity link and obtain

$$P_i(y_{if}|\boldsymbol{z}, \boldsymbol{w}) = \exp\left[\frac{y_{if}(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z}) - (b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z})^2/2}{\phi_i} - y_{if}^2/(2\phi_i) - \log(2\pi\phi_i)/2\right]. \tag{5}$$

If the continuous data are not normally distributed, such as positively skewed response times, it is common to apply a log-transformation (e.g., De Boeck & Scalise, 2019; van der Linden, 2006, 2007; Wang, Xu, & Shang, 2018) before applying Equation 5 in the field of educational and psychological measurement.

3. *Count responses.* We consider Poisson and negative-binomial distributions with a log link function for count data. In the former case, we have

$$P_i(y_{if} = c|\boldsymbol{z}, \boldsymbol{w}) = \frac{\lambda_i^c}{c!}\exp(-\lambda_i), \tag{6}$$

where $\lambda_i = \exp(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z})$. With a negative binomial distribution, we have the conditional probability mass function (Niku et al., 2017)

$$P_i(y_{if} = c|\boldsymbol{z}, \boldsymbol{w}) = \frac{\Gamma\left(c + \frac{1}{\phi_i}\right)}{c!\Gamma\left(\frac{1}{\phi_i}\right)}\left(\frac{\exp(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z})}{\frac{1}{\phi_i} + \exp(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z})}\right)^c\left(\frac{1}{1 + \phi_i\exp(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z})}\right)^{\frac{1}{\phi_i}}, \tag{7}$$

where $\Gamma(\cdot)$ indicates the gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$.

### Laplace approximations for GLLVMs

As mentioned above, Equation 3 does not have an explicit solution, requiring approximation methods for parameter estimation. In this article, we utilize Laplace approximations to approximate the integrals in the likelihood function. We define $h_f(\boldsymbol{z}) = -\log P(\boldsymbol{y}_f|\boldsymbol{z})\psi(\boldsymbol{z};\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\hat{h} = h_f(\hat{\boldsymbol{z}}_f)$, and $\boldsymbol{H}_f = \frac{\partial^2 \hat{h}}{\partial z \partial z'}$, where $\hat{\boldsymbol{z}}_f$ represents the posterior modes of the latent scores of individual $f \in 1, \cdots, N$. The second-order Laplace approximation of the marginal log-likelihood for an individual $f$ can then be written as (Shun, 1997)

$$\tilde{l}_f^{\text{Lap2}}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{P}{2}\log(2\pi) - \frac{1}{2}|\boldsymbol{H}_f| - \hat{h} + \log(1+\epsilon_f), \tag{8}$$

with

$$\epsilon_f = -\frac{1}{2}\left[\frac{1}{4}\sum_{jklm}^{P}\frac{\partial^4 \hat{h}}{\partial z_j \partial z_k \partial z_l \partial z_m} b_{jl}b_{km} - \frac{1}{4}\sum_{jklrst}^{P}\frac{\partial^3 \hat{h}}{\partial z_j \partial z_k \partial z_l}\frac{\partial^3 \hat{h}}{\partial z_r \partial z_s \partial z_t} b_{jr}b_{kl}b_{st} \right. \tag{9}$$
$$\left. - \frac{1}{6}\sum_{jklrst}^{P}\frac{\partial^3 \hat{h}}{\partial z_j \partial z_k \partial z_l}\frac{\partial^3 \hat{h}}{\partial z_r \partial z_s \partial z_t}\frac{1}{6} b_{jr}b_{ks}b_{lt}\right],$$

where $b_{jk}$ represents the entry of row $j$ and column $k$ in $\boldsymbol{H}_f^{-1}$. By setting $\epsilon_f = 0$ in Equation 8, the second-order Laplace approximation reduces to the first-order Laplace approximation. To efficiently compute $\epsilon_f$, it is necessary to consider the particular model structure used and identify unique and zero entries of $\epsilon_f$. Readers are directed to Andersson et al. (2023) for details of the filtering procedure used to compute $\epsilon_f$. We utilize the same estimation approach that Jin and Andersson (2020) and Andersson et al. (2023) proposed for categorical observed variables and extend it to support continuous and count data measurement models, where the derivatives in Equation 8 and its gradient are derived analytically. Each entry $\theta \in \boldsymbol{\theta}$ of the gradient is given by

$$\nabla_f^\theta = \frac{\partial l_f^{\text{Lap2}}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta} + \frac{\partial \hat{\boldsymbol{z}}_f}{\partial \theta}\frac{\partial l_f^{\text{Lap2}}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{z}}\Bigg|^{\boldsymbol{z}=\hat{\boldsymbol{z}}_f}, \tag{10}$$

where the second term is needed to account for the dependence between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{z}}_f$. The needed derivatives (up to the fifth order) are presented in Appendix B. A quasi-Newton method using the BFGS algorithm is utilized to maximize the approximated marginal log-likelihood function.

## Motivating example

In this section, we provide an example based on the computer-based assessment of mathematics (CBAM) in PISA 2012, which aims to assess 15-year-old students' mathematical literacy and reflects the importance of using digital tools to solve mathematics tasks (Peña-López et al., 2012). Students can, for example, rotate representations of 3D objects and draw points and lines to facilitate their thinking processes. The full CBAM instrument consists of 41 items from 15 units and the items are organized into four clusters. Each student was given two clusters with 40 minutes total testing time (Peña-López et al., 2012). PISA released three units out of 15 and the data are available on the website of the Organisation for Economic Co-operation and Development. We chose unit CM015 (*CD Production*) to illustrate the practical use of the proposed method and to guide our simulation design. CM015 presents an interactive graph and a price calculator and asks participants to enter the number of copies to discover its relationship with the cost of copying CDs using duplication and replication methods. Three items were included in CM015 with one multiple-choice and two constructed-response items. As an example, we used the Australian data set because it had the largest sample size ($N = 1824$) participating in this unit.

Three indicators were extracted: task scores, response time, and the number of actions. We pre-processed the data by a) log-transforming and centering the response time to deal with its positively skewed distribution and b) excluding outliers in terms of response times and the number of actions that were beyond the range from $Q1 - 3 \times IQR$ to $Q3 + 3 \times IQR$, where Q1, Q3, and IQR represent the first quantile, the third quantile,

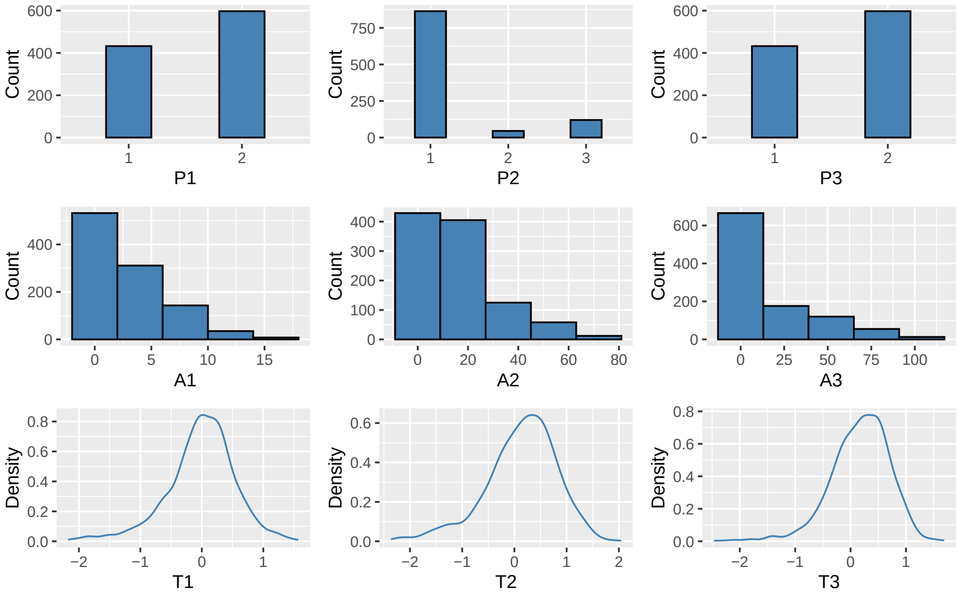


**Figure 1**

*Summary of observed indicators. P1 - P3, A1 - A3, and T1 - T3 represent scored responses, the number of actions, and transformed response times of Items 1 to 3, respectively.*

and the interquartile range, respectively. After excluding these outliers and missing values, 1029 respondents remained and were used for the following analysis. A summary of the three indicators is presented in Figure 1. We then separately applied unidimensional measurement models to responses, the number of actions, and response times using Lap2, and used the parameter estimates as a reference for the following simulation studies.

Next, we combined the three indicators in a single model using GLLVMs with the residual correlations of indicators from the same item considered (ModRes, see Figure 5) or not (ModInd, see Figure 2). Equality constraints were added to residual factor loadings from the same item for simplification. In total, we estimated 2 (model structure: ModInd or ModRes) × 2 (count type: Poisson or negative-binomial) × 2 (algorithms: Lap1 or Lap2) = 8 models. Lap1 failed to converge with ModRes and either type of count data

**Table 1**

*Model fit of generalized linear latent variable models using the empirical data.*

| Count data model | Method | ModInd | | ModRes | |
|---|---|---|---|---|---|
| | | BIC | SRMSR | BIC | SRMSR |
| Poisson | Lap1 | 32925 | 0.127 | - | - |
| Poisson | Lap2 | 32897 | 0.126 | 28325 | 0.088 |
| Negative-binomial | Lap1 | 28372 | 0.086 | - | - |
| Negative-binomial | Lap2 | 28353 | 0.086 | 27923 | 0.070 |

*Notes.* Lap1 did not converge with ModRes. Lap1 = first-order Laplace, Lap2 = second-order Laplace, BIC = Bayesian information criterion, SRMSR = standardized root mean square residuals.

model, whereas Lap2 achieved convergence in all models. The model fit is presented in Table 1. In the case of ModInd, a negative-binomial distribution fit better than the Poisson distribution. The BIC of ModRes was smaller than that of ModInd. Given that ModRes with a negative-binomial distribution using Lap2 had the lowest BIC, we concluded that this model best represented the observed data and present item parameter estimates in Appendix A1.

## Simulation study

We conducted two simulation studies to assess the performance of Laplace approximations in the context of mixed data using newly developed code written in C++ and R 4.1.2 (R Core Team, 2021). In Simulation 1, we considered three-dimensional GLLVMs with three types of indicators: ordinal, continuous, and count data. In Simulation 2, we also considered residual correlations between indicators from the same task.

We evaluated the performance of the proposed methods in terms of convergence rate, estimation time, and the recovery of model parameters. Convergence was determined by satisfying two criteria: a) the algorithm stopped before 500 iterations, and b) the approximated observed information matrix was positive definite. In addition to excluding non-converged replications, we also excluded unstable replications with extreme outliers which were defined as replications that had estimates with absolute bias larger than 5. Regarding parameter recovery, we computed the absolute bias and the mean squared error (MSE) to assess the accuracy and precision of parameter estimates via

$$|\text{bias}|_\theta = |\sum_{r=1}^{R}(\hat{\theta}^r - \theta)/R|, \tag{11}$$

and

$$\text{MSE}_\theta = \sum_{r=1}^{R}(\hat{\theta}^r - \theta)^2/R, \tag{12}$$

where $\theta$ and $\hat{\theta}^r$ represent indicates the true value and estimate of a parameter in the replication $r \in 1, \cdots, R$, respectively.

## First simulation study

### *Simulation 1 design*

In Simulation 1, we considered three correlated latent variables with ordinal, continuous, and count data as indicators, respectively. We illustrate the model in Figure 2. Three experimental factors were manipulated: a) the distribution of the count data model (Poisson or negative-binomial distributions), b) the number of items (3 or 6), and c) the covariance between the latent variables (small and large). This resulted in $2 \times 2 \times 2 = 8$ conditions. We used 1000 replications for each condition. To determine the ranges of the simulated parameters, we made use of the result from the motivating example and the PISA 2018 item parameter pool. Specifically, the item pool provides the estimates of item parameters in terms of task scores, and we used the 10% quantile and 90% quantile as the range of the item parameters for the ordinal data model. For the models corresponding to response time and number of actions, we generated the item parameters based on the above motivating example. The distributions from which we generated the true parameters are presented in Table 2. Latent variables were randomly simulated from a multivariate normal distribution with a zero mean vector and variances equal to one. The sample size was fixed at 1000. The observed data were then generated based on Equations 4-7. With the datasets generated, we estimated the measurement models for each outcome variable and if convergence was achieved, the parameter estimates were used as the starting values to estimate the three-dimensional model with the first- and second-order Laplace approximation methods.

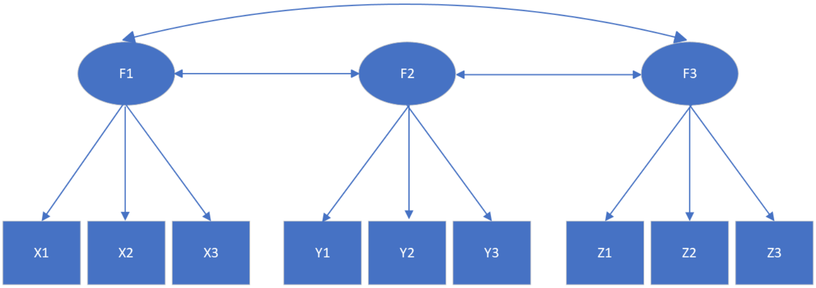


**Figure 2**

*Model illustration of three-dimensional GLLVMs. X, Y, and Z indicate three different types of indicators. F1-F3 indicate latent variables.*

### ***Simulation 1 results***

The convergence rates and estimation times of each algorithm are presented in Table 3. It indicates that Lap2 outperformed Lap1 under the conditions with three items and both methods reached 100% convergence as the number of items increased. Among the converged and stable replications, both methods accomplished the estimation procedure within 15 seconds on average. As expected, Lap1 consumed less time than Lap2 in all conditions, but the difference was minor. Regarding the experimental factors, Table 3 suggests that increasing the number of items or using a negative-binomial distribution took

**Table 2**

*Distributions of true parameters.*

| Data type | Parameter | Distribution |
|---|---|---|
| *Ordinal data* | $a$ slope parameter | $U(0.74, 1.69)$ |
| | $b1$ threshold parameter | $U(0.2, 1.25)$ |
| | $b2$ threshold parameter | $U(-1.25, -0.2)$ |
| *Continuous data* | $a$ slope parameter | $U(0.4, 0.8)$ |
| | $b$ intercept parameter | $U(-0.2, 0.2)$ |
| | $\phi$ scale parameter | $U(0.1, 0.3)$ |
| *Count data* | $a$ slope parameter | $U(0.4, 0.8)$ |
| | $b$ intercept parameter | $U(1, 3)$ |
| | $\phi$ scale parameter (negative binomial) | $U(0.5, 1)$ |
| *Covariance parameters* | $\rho$ covariance: small | $U(0.2, 0.4)$ |
| | $\rho$ covariance: large | $U(0.6, 0.8)$ |

longer time in estimation.

Next, we summarize the recovery of parameters. Overall, the parameters showed small absolute bias and MSE, indicating that the methods recovered the true parameters accurately and precisely. We plot the absolute bias and MSE of model parameters with test length three in Figure 3. Figure 3 indicated that: a) Lap2 produced less bias in estimating the parameters regarding ordinal data, count data, and covariance than Lap1 did, and b) the absolute biases of the item parameters of the continuous data using Lap1 and Lap2 were visually indistinguishable. We also present the MSE of the estimates in Figure 4. In general, the values of MSE were small and similar for both Lap1 and Lap2. However, Lap1 estimated the slope parameters of ordinal data more precisely when the correlation between the latent variables was small. The six-item conditions showed similar

**Table 3**

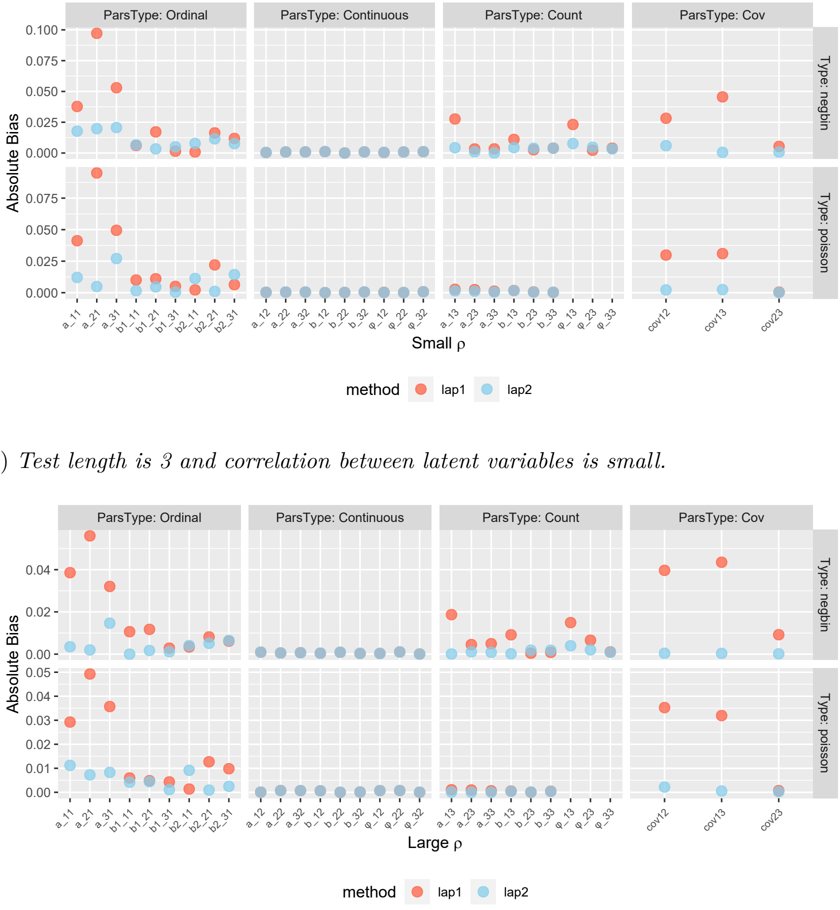
*Convergence rate and timing (seconds) of Lap1 and Lap2 in Simulation 1.*

| Con | Type | #Item | $\rho$ | Convergence rate | | Timing | |
|---|---|---|---|---|---|---|---|
| | | | | Lap1 | Lap2 | Lap1 | Lap2 |
| 1 | Pois | 3 | small | 81.5% | 86.4% | 9.15 | 10.58 |
| 2 | Pois | 3 | large | 84.2% | 84.6% | 8.99 | 10.67 |
| 3 | Pois | 6 | small | 100% | 100% | 15.29 | 17.10 |
| 4 | Pois | 6 | large | 100% | 100% | 14.27 | 15.96 |
| 5 | Negbin | 3 | small | 82.9% | 86.2% | 9.25 | 10.79 |
| 6 | Negbin | 3 | large | 80.7% | 84.8% | 9.37 | 11.02 |
| 7 | Negbin | 6 | small | 100% | 100% | 17.51 | 20.30 |
| 8 | Negbin | 6 | large | 100% | 100% | 16.06 | 18.43 |
| Overall | | | | 91.2% | 92.8% | 12.49 | 14.36 |

*Notes.* Lap1 = first-order Laplace, Lap2 = second-order Laplace, Pois = Poisson, Negbin = negative-binomial, $\rho$ = covariance of latent variables.

patterns and can be found in the supplementary material[1]. To illustrate the differences between the methods, we considered the cases where the absolute average bias was larger than .01 (i.e., slope and threshold parameters for ordinal data and slope and scale parameters for negative-binomial distributed data) and present the results under each simulating factor in Table 4. It suggested that a) increasing the number of items or the covariance between the latent variables improved the accuracy of item parameter estimates, especially for Lap1, and b) when estimating the covariance parameter, Lap2 was less influenced by the simulating factors and recovered covariance satisfactorily under all conditions, whereas Lap1 showed a relatively large bias.

[1] Link: `https://osf.io/nec8m/?view_only=fc93a3e633ea47eba597357722fe8c83`

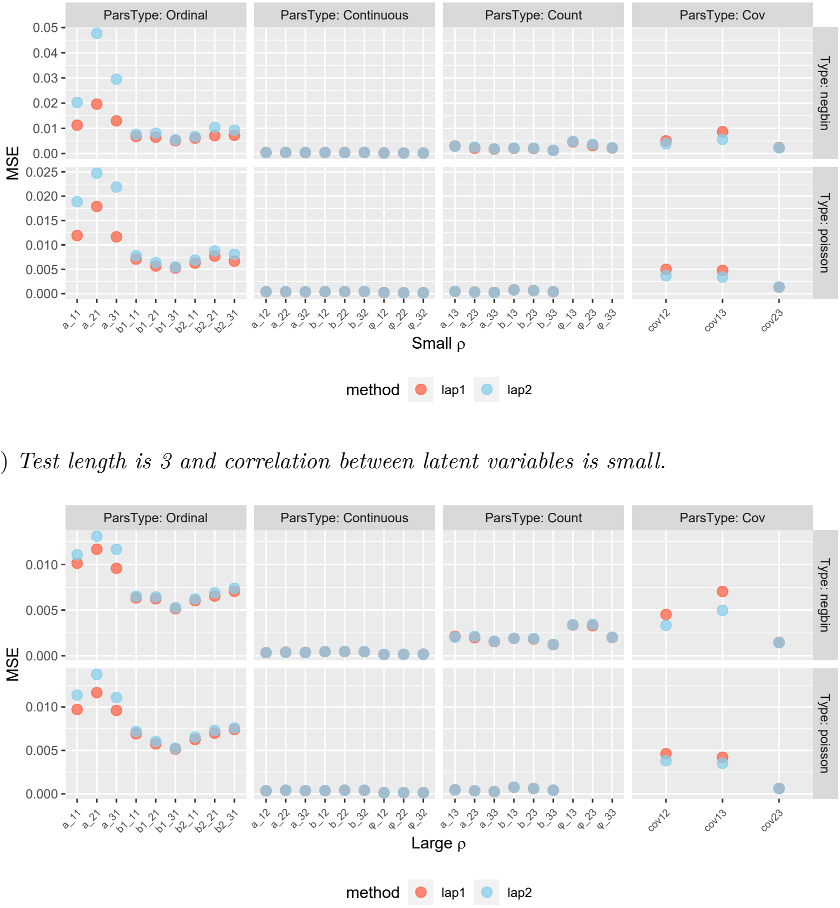


(a) *Test length is 3 and correlation between latent variables is small.*

(b) *Test length is 3 and correlation between latent variables is large.*

**Figure 3**

*Absolute bias of item parameters and covariances in Simulation 1 when test length is 3, where a, b, and $\phi$ represent the slope, intercept, and scale parameters, b1 and b2 are thresholds, and a_ip represents the slope parameter of latent variable $F_p$ on item $i$.*



(a) *Test length is 3 and correlation between latent variables is small.*

(b) *Test length is 3 and correlation between latent variables is large.*

**Figure 4**

*MSE of item parameters and covariances in Simulation 1 when test length is 3, where a, b, and $\phi$ represent the slope, intercept, and scale parameters, b1 and b2 are thresholds, and a_ip represents the slope parameter of latent variable $F_p$ on item i.*

**Table 4**

*Absolute bias of more biased item parameters under simulating factors in Simulation 1.*

| Factor | Level | Method | Ordinal model | | Negbin model | | Covariance |
|---|---|---|---|---|---|---|---|
| | | | Slope | Thresholds | Slope | Scale | |
| #Item | 3 | Lap1 | 0.0512 | 0.0080 | 0.0104 | 0.0087 | 0.0251 |
| | 6 | Lap1 | 0.0286 | 0.0064 | 0.0039 | 0.0016 | 0.0096 |
| | 3 | Lap2 | 0.0124 | 0.0048 | 0.0012 | 0.0037 | 0.0013 |
| | 6 | Lap2 | 0.0070 | 0.0032 | 0.0011 | 0.0025 | 0.0008 |
| $\rho$ | small | Lap1 | 0.0447 | 0.0088 | 0.0069 | 0.0049 | 0.0158 |
| | large | Lap1 | 0.0276 | 0.0051 | 0.0052 | 0.0030 | 0.0189 |
| | small | Lap2 | 0.0114 | 0.0044 | 0.0015 | 0.0035 | 0.0015 |
| | large | Lap2 | 0.0062 | 0.0031 | 0.0008 | 0.0024 | 0.0007 |
| Overall | | Lap1 | 0.0362 | 0.0069 | 0.0061 | 0.0039 | 0.0173 |
| | | Lap2 | 0.0088 | 0.0037 | 0.0012 | 0.0029 | 0.0011 |

*Notes.* Lap1 = first-order Laplace, Lap2 = second-order Laplace, Negbin = negative-binomial, $\rho$ = covariance of latent variables.

## Second simulation study

### *Simulation 2 design*

In Simulation 2, we considered residual correlations of indicators from the same item. Compared to Simulation 1, we added three residual latent variables - R1, R2, and R3 - to capture the item-specific effect for single items (Figure 5). We specified that the residual latent variables impose the same effect on the indicators, namely, we set equal residual factor loading across indicators from the same item (e.g., equal residual factor loadings for X1, Y1, and Z1). Accordingly, we added the magnitude of the residual factor loadings (small or large) to the simulation design. Specifically, large residual factor loadings were generated from $U(0.4, 0.8)$, the same distribution as the slope parameter for

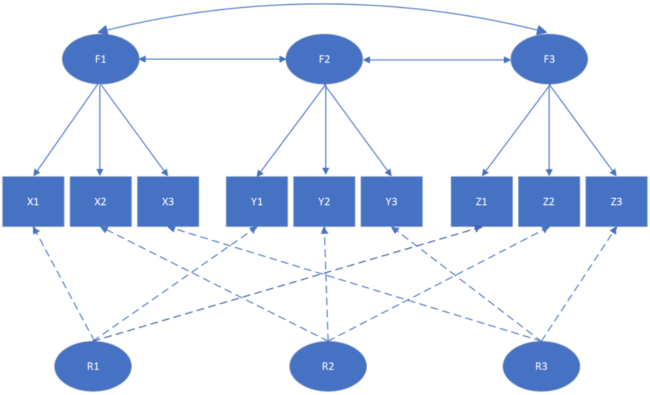


**Figure 5**

*Model illustration including residual latent variables, where X, Y, and Z indicate three different types of indicators, F1-F3 are latent variables measured by the X-, Y-, and Z-variables, respectively, and R1-R3 are residual latent variables.*

the continuous data and count data, whereas the small residual factor loading was set to half of the large one. In sum, Simulation 2 resulted in 2 (Poisson or Negative-binominal distribution) × 2 (3 or 6 items) × 2 (small or large covariance) × 2 (small or large residual factor loading) = 16 conditions and we generated 1000 datasets under each condition. Both Lap1 and Lap2 were applied to analyze the datasets.

### ***Simulation 2 results***

We now summarize the result from Simulation 2. In line with Simulation 1, Lap2 reached a higher average convergence rate (98.8%) than Lap1 did (93.9%) (Table 5). The main difference occurred when the number of items was three and the residual factor loading was small. However, Lap2 cost much longer time than Lap1, especially as the number of items increased. Regarding parameter recovery, Lap2 estimated item parameters more accurately than Lap1, especially for ordinal data and count data with a negative-binomial distribution. This conclusion was in line with Simulation 1. However, note that Lap1 and Lap2 differed a bit in the recovery of item parameters for continuous data in Simulation 2 because of the addition of residual correlations. As for the added residual factor loadings, we present their absolute biases in Figure 6. The estimation of residual factor loadings was both accurate (average absolute bias less than .008) and precise (average MSE less than .005).

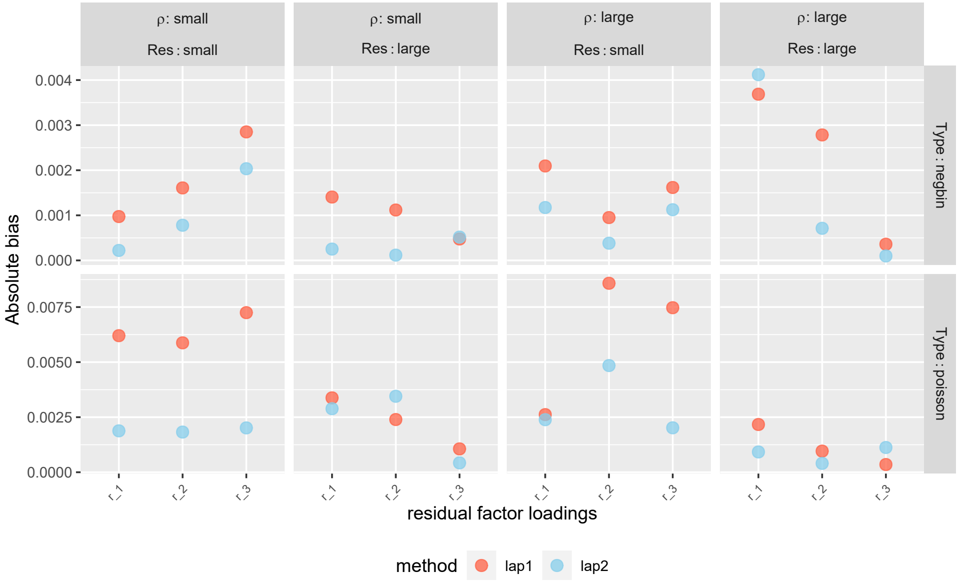


**Figure 6**

*Absolute bias of residual factor loadings in Simulation 2 when the number of items is 3.*

As in Simulation 1, we inspected the estimates with biases larger than 0.01 more thoroughly. It turned out that these biases primarily referred to the slope and threshold parameter of ordinal data, the slope and scale parameter of count data, and the covariance of latent variables. We organized the above-mentioned estimates by the experimental factors in Table 6. With the number of items increasing, both Lap1 and Lap2 estimated item parameters and the covariance more accurately according to the great decrease of average absolute bias in Table 6. Regarding the correlation of latent variables, decreasing

**Table 5**

*Convergence rates and timing (seconds) of Lap1 and Lap2 in simulation 2.*

| Type | #Item | $\rho$ | residuals | Convergence rate | | Timing | |
|---|---|---|---|---|---|---|---|
| | | | | Lap1 | Lap2 | Lap1 | Lap2 |
| Pois | 3 | small | small | 87.6% | 97.1% | 13.7 | 56.5 |
| Pois | 3 | small | large | 93.3% | 97.6% | 15.1 | 63.5 |
| Pois | 3 | large | small | 84.7% | 95.5% | 13.2 | 52.7 |
| Pois | 3 | large | large | 91.1% | 99.8% | 15.3 | 59.6 |
| Negbin | 3 | small | small | 82.9% | 97.7% | 14.0 | 61.8 |
| Negbin | 3 | small | large | 92.0% | 98.2% | 15.2 | 69.1 |
| Negbin | 3 | large | small | 81.7% | 95.2% | 14.4 | 60.8 |
| Negbin | 3 | large | large | 90.5% | 99.8% | 15.5 | 66.0 |
| Pois | 6 | small | small | 100% | 100% | 40.1 | 351.4 |
| Pois | 6 | small | large | 99.6% | 99.7% | 47.4 | 410.8 |
| Pois | 6 | large | small | 100% | 100% | 38.0 | 333.9 |
| Pois | 6 | large | large | 99.2% | 99.7% | 48.7 | 423.0 |
| Negbin | 6 | small | small | 100% | 100% | 44.8 | 413.7 |
| Negbin | 6 | small | large | 100% | 100% | 44.7 | 416.8 |
| Negbin | 6 | large | small | 100% | 100% | 43.9 | 405.9 |
| Negbin | 6 | large | large | 100% | 100% | 45.0 | 415.5 |
| Overall | | | | 93.9% | 98.8% | 29.32 | 228.81 |

*Notes.* Lap1 = first-order Laplace, Lap2 = second-order Laplace, Pois = Poisson, Negbin = negative-binomial, $\rho$ = covariance of latent variables.

the correlation almost doubled the absolute bias for the slope and threshold parameters in ordinal data for both Lap1 and Lap2. However, such an effect became minor in estimating the slope and scale parameters in negative-binomial data. When the true correlation increased, the average absolute bias of the covariance estimate increased using Lap1 but decreased slightly using Lap2. The magnitude of residual loading had a small influence on parameter recovery.

**Table 6**

*Absolute bias of more biased item parameters under simulating factors in Simulation 2.*

| Factor | Level | Method | Ordinal model | | Negbin model | | Covariance |
|---|---|---|---|---|---|---|---|
| | | | Slope | Thresholds | Slope | Scale | |
| #Item | 3 | Lap1 | 0.0626 | 0.0089 | 0.0171 | 0.0161 | 0.0317 |
| | 6 | Lap1 | 0.0319 | 0.0067 | 0.0075 | 0.0044 | 0.0110 |
| | 3 | Lap2 | 0.0247 | 0.0090 | 0.0025 | 0.0039 | 0.0037 |
| | 6 | Lap2 | 0.0075 | 0.0033 | 0.0013 | 0.0027 | 0.0013 |
| $\rho$ | small | Lap1 | 0.0509 | 0.0091 | 0.0119 | 0.0085 | 0.0184 |
| | large | Lap1 | 0.0333 | 0.0058 | 0.0095 | 0.0081 | 0.0243 |
| | small | Lap2 | 0.0185 | 0.0070 | 0.0017 | 0.0032 | 0.0038 |
| | large | Lap2 | 0.0080 | 0.0034 | 0.0018 | 0.0030 | 0.0012 |
| *Res* | small | Lap1 | 0.0402 | 0.0075 | 0.0098 | 0.0066 | 0.0213 |
| | large | Lap1 | 0.0440 | 0.0074 | 0.0116 | 0.0100 | 0.0214 |
| | small | Lap2 | 0.0110 | 0.0044 | 0.0017 | 0.0030 | 0.0017 |
| | large | Lap2 | 0.0155 | 0.0060 | 0.0017 | 0.0032 | 0.0033 |
| Overall | | Lap1 | 0.0421 | 0.0074 | 0.0107 | 0.0083 | 0.0214 |
| | | Lap2 | 0.0132 | 0.0052 | 0.0017 | 0.0031 | 0.0025 |

*Notes.* Lap1 = first-order Laplace, Lap2 = second-order Laplace, Negbin = negative-binomial, $\rho$ = covariance of latent variables.

## Discussion

The advent of complex measurement tools has facilitated research by providing more detailed information of the response process. As a result, the data often consists of different types. In this paper, we implemented first- and second-order Laplace approximations to jointly model a mixture of ordinal, continuous, and count data within the framework of GLLVMs. An empirical study demonstrated the usage of the proposed methods in practice and two simulation studies were conducted to examine the performance of both algorithms in the scenario of computer-based assessment with process indicators and performance data. The results indicated that Lap2 had a higher convergence rate and better parameter recovery compared to Lap1. However, Lap2 took longer to estimate, especially with complex models that incorporated residual factors.

The experimental factors impacted the results in the following ways. First, test length had a significantly positive influence on convergence and the recovery of item parameters. As the number of items increases, both Lap1 and Lap2 approximated the marginal log-likelihood better and the error of the estimators decreased (Huber et al., 2004). Moreover, higher-order Laplace approximations have a faster rate of error decrease (Andersson & Xin, 2021) which means that fewer indicators are needed for accurate estimation. Second, the magnitude of covariance between latent variables had a positive effect on the estimation. Third, the magnitude of residual factor loadings had some influence on the convergence and estimation time but a minor effect on parameter recovery. Larger residual factor loadings implied stronger item-specific effects, which should be considered in the model specification. In the empirical study, we found that considering the residual correlation improved the model fit and the residual factor loadings were not negligible (see Appendix A1). In this study, we imposed equality restrictions on residual factor loadings from the same item. If there are prior hypotheses about the residual factor loadings, it is flexible to take them into account and specify such a hypothesized model as long as it is identifiable.

## Contributions and limitations

The study makes significant contributions in several ways. First, we derived the second-order Laplace approximation likelihood form of Poisson and negative-binomial distributions for count data, which extended existing research that only included first-order Laplace approximations for count data (Niku et al., 2017). Employing Lap2 can improve the estimation accuracy in terms of the slope and scale parameters compared to using Lap1 for data with a negative-binomial distribution. Second, the current study provided a fast yet accurate solution for a combination of count data, continuous data, and ordinal data within the framework of GLLVMs. Compared to a Bayesian or quadrature approach, Laplace approximations greatly increase the computational efficiency in high-dimensional GLLVMs (Huber et al., 2004). Our research considered different types of observed variables and potential residual correlations between the indicators in a single model. This extended a) the study of Niku et al. (2017) by considering different types of indicators at the same time and b) the research related to joint modeling of responses and response times within a hierarchical framework (van der Linden, 2007) by incorporating count data. Third, compared to Andersson et al. (2023) that only considered categorical data, we compared the first- and second-order Laplace approximation in the case of GLLVMs with a mixture of ordinal, count, and continuous indicators, which advanced our knowledge of the performance of both algorithms in the mixed-data situation with different test lengths, correlations of latent variables, and magnitudes of residual factor loadings.

On the other hand, some limitations of the paper should be noted. The first limitation is shared with the approach of Lap2 - it is required to compute up to fifth-order derivatives based on the selected distributions of observed indicators. This means that substantial derivations are necessary to support additional distributions. In the current study, only continuous, count, and ordinal data were considered. However, it is feasible to consider other types/distributions of indicators, which is a potential direction for future studies. Second, because of the focus on likelihood-based estimation we only compared the first- and second-order Laplace approximations but did not consider other approaches such as Bayesian methods. We could also not compare our results to regular or adaptive Gauss-Hermite quadrature due to the computational expense of these methods.

### Practical suggestions

In this study, we used process data and performance data from computer-based assessments to demonstrate the application of the proposed method. However, the method has the potential to be applied to broader areas beyond psychological and educational assessment. For example, a combination of different data types often occurs in ecological data such as species counts and biomass in biology (Niku et al., 2017) and patient data relevant to symptoms such as presence/absence, frequency, and scale scores in health (Daniels & Normand, 2006).

For practitioners dealing with a mixture of different types of data, we offer some suggestions. First, when there are more than two latent variables, Laplace approximations have a great advantage over numerical quadrature or Bayesian approaches in terms of computational efficiency. Within Laplace approximations, Lap1 is faster than Lap2 and the efficiency advantage increases with the dimension of latent variables and the complexity of model structures. For example, the difference between the average time for estimating three-dimensional models (Simulation 1) was two seconds, while the value increased to 200 seconds for six-dimensional models with residual correlations considered (Simulation 2). Second, when the number of items is small such as three items per dimension, it is suggested to use Lap2 because it has a higher convergence rate compared to Lap1. It is also worth noting that Lap2 significantly improves the non-convergence problem of Lap1 when the observed variables are dichotomous (Andersson & Xin, 2021). Third, starting values have a large impact on the estimation of GLLVMs. This is because the observed likelihood can be multimodal when GLLVMs have a complex mean and latent variable structure (Niku et al., 2019). If researchers or practitioners have prior knowledge of the estimates based on existing literature or studies, it is possible to make use of that information. If no prior knowledge is available, it is possible to make use of the data provided to determine the starting values (Niku et al., 2019). In our simulation studies, we first fit unidimensional measurement models and obtained the estimates as starting values of item parameters for three- or six-dimensional models if the unidimensional models converged. This greatly reduced the estimation time and increased the convergence rate.

# References

Andersson, B., & Jin, S. (2022). *lamle: Maximum likelihood estimation of latent variable models using adaptive quadrature and laplace approximations.* `https://github.com/bjoernhandersson/lamlepub/releases/tag/v0.1.2-alpha`.

Andersson, B., Jin, S., & Zhang, M. (2023). Fast estimation of multiple group generalized linear latent variable models for categorical observed variables. *Computational Statistics & Data Analysis*, *182*, 107710.

Andersson, B., & Xin, T. (2021). Estimation of latent regression item response theory models using a second-order laplace approximation. *Journal of Educational and Behavioral Statistics*, *46*(2), 244–265.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach.* John Wiley & Sons.

Daniels, M. J., & Normand, S.-L. T. (2006). Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*, *7*(1), 1–15.

De Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology*, *10*, 1280.

De Leon, A. R., & Chough, K. C. (2013). *Analysis of mixed data: Methods & applications.* CRC Press.

Huber, P., Ronchetti, E., & Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(4), 893–908.

Jin, S., & Andersson, B. (2020). A note on the accuracy of adaptive gauss–hermite quadrature. *Biometrika*, *107*, 737–744.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, *52*, 5066–5074.

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2021). Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness. *Journal of Applied Psychology*.

Man, K., & Harring, J. R. (2022). Detecting preknowledge cheating via innovative measures: A mixture hierarchical model for jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement*, 00131644221136142.

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*(1), 1–27.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, *49*(2), 313–334.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*(3), 391–411.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muthén, L. K., & Muthén, B. (2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide*. Muthén & Muthén.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K., Taskinen, S., & Warton, D. I. (2019). Efficient estimation of generalized linear latent variable models. *PloS ONE*, *14*(5), e0216129.

Niku, J., Warton, D. I., Hui, F. K., & Taskinen, S. (2017). Generalized linear latent variable models for multivariate count and biomass data in ecology. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(4), 498–522.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, *47*, 337–347.

Peña-López, I., et al. (2012). *Pisa 2012 assessment and analytical framework. mathematics, reading, science, problem solving and financial literacy.* OECD Publishing.

Qiao, X., Jiao, H., & He, Q. (2022). Multiple-group joint modeling of item responses, response times, and action counts with the conway-maxwell-poisson distribution. *Journal of Educational Measurement.*

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*(1), 1–21.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190.

Shun, Z. (1997). Another look at the salamander mating data: A modified laplace approximation approach. *Journal of the American Statistical Association*, *92*(437), 341–349.

Steinfeld, N. (2016). "I agree to the terms and conditions": (How) do users read privacy policies online? An eye-tracking experiment. *Computers in Human Behavior*, *55*, 992–1000.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*, *55*(3), 425–453.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*, 223–254.

# Appendix A

## The results of the final model in the motivating example

**Table A1**

*Parameter estimates (standard error) of the final model.*

| | Indicator | Slope | Intercept 1 | Intercept 2 | Scale |
|---|---|---|---|---|---|
| Performance | P1 | 1.361 (0.125) | 0.441 (0.088) | - | - |
| | P2 | 2.695 (0.312) | -3.346 (0.317) | -3.987 (0.355) | - |
| | P3 | 3.093 (0.288) | 0.120 (0.139) | -3.042 (0.264) | - |
| Action | A1 | 0.515 (0.087) | 1.101 (0.029) | - | 0.080 (0.020) |
| | A2 | 0.421 (0.025) | 2.679 (0.025) | - | 0.209 (0.017) |
| | A3 | 1.046 (0.039) | 2.315 (0.044) | - | 0.272 (0.038) |
| Time | T1 | 0.288 (0.019) | -0.018 (0.017) | - | 0.108 (0.012) |
| | T2 | 0.481 (0.023) | 0.132 (0.021) | - | 0.079 (0.015) |
| | T3 | 0.388 (0.018) | 0.170 (0.017) | - | 0.009 (0.013) |
| Residual | Item 1 | 0.355 (0.018) | - | - | - |
| | Item 2 | 0.382 (0.018) | - | - | - |
| | Item 3 | 0.354 (0.023) | - | - | - |

*Note.* P2 and P3 have three categories and thus have two intercept parameters. The correlations between the latent variables were 0.876, 0.658, and 0.584.

# Appendix B

## Derivatives

The needed derivatives in Equation 8 are presented as follows. Let $h_i = -\log P_{iy_{if}}$ and define $\mathbf{1}()$ as an indicator function.

### Ordinal data

We adopt the generalized partial credit model (GPCM) for ordinal responses. Let $P_{ic}$ represent $P_i(y_{if} = c|\boldsymbol{z}, \boldsymbol{w})$. The derivatives of $h_i$ with respect to $\boldsymbol{z}$ are

$$\frac{\partial h_i}{\partial z_j} = -a_{ij}\left(y_{if} - \sum_{c=1}^{m_i} cP_{ic}\right),$$

$$\frac{\partial^2 h_i}{\partial z_j \partial z_k} = a_{ij}\sum_{c=1}^{m_i} c\frac{\partial P_{ic}}{\partial z_k},$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial z_l} = a_{ij}\sum_{c=1}^{m_i} c\frac{\partial^2 P_{ic}}{\partial z_k \partial z_l},$$

$$\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m} = a_{ij}\sum_{c=1}^{m_i} c\frac{\partial^3 P_{ic}}{\partial z_k \partial z_l \partial z_m},$$

and

$$\frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial z_n} = a_{ij}\sum_{c=1}^{m_i} c\frac{\partial^4 P_{ic}}{\partial z_k \partial z_l \partial z_m \partial z_n}.$$

The derivatives of $h_i$ with respect to $u \in \{a_i, b_{i2}, \ldots b_{im_i}\}$ are

$$\frac{\partial h_i}{\partial u} = -\frac{\frac{\partial P_{iy_f}}{\partial u}}{P_{iy_f}},$$

$$\frac{\partial^2 h_i}{\partial z_j \partial u} = -\mathbf{1}(u = a_{ij})\left(y_{if} - \sum_{c=1}^{m_i} c\frac{\partial P_{ic}}{\partial u}\right) + a_{ij}\sum_{c=1}^{m_i} c\frac{\partial P_{ic}}{\partial u},$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial u} = \mathbf{1}(u = a_{ij})\sum_{c=1}^{m_i} c\frac{\partial P_{ic}}{\partial z_k} + a_{ij}\sum_{c=1}^{m_i} c\frac{\partial^2 P_{ic}}{\partial z_k \partial u},$$

$$\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial u} = \mathbf{1}(u = a_{ij}) \sum_{c=1}^{m_i} c \frac{\partial^2 P_{ic}}{\partial z_k \partial z_l} + a_{ij} \sum_{c=1}^{m_i} c \frac{\partial^3 P_{ic}}{\partial z_k \partial z_l \partial u},$$

and

$$\frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial u} = \mathbf{1}(u = a_{ij}) \sum_{c=1}^{m_i} c \frac{\partial^3 P_{ic}}{\partial z_k \partial z_l \partial z_m} + a_{ij} \sum_{c=1}^{m_i} c \frac{\partial^4 P_{ic}}{\partial z_k \partial z_l \partial z_m \partial u}.$$

The derivatives of $P_{ic}$ with respect to $\boldsymbol{z}$ in the above equations are

$$\frac{\partial P_{ic}}{\partial z_k} = P_{ic} a_{ik} \left[ c - \sum_{c'=1}^{m_i} c' P_{ic'} \right],$$

$$\frac{\partial^2 P_{ic}}{\partial z_k \partial z_l} = \frac{\partial P_{ic}}{\partial z_l} a_{ik} \left[ c - \sum_{c'=1}^{m_i} c P_{ic} \right] - P_{ic} a_{ik} \sum_{c=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_l},$$

$$\begin{aligned} \frac{\partial^3 P_{ic}}{\partial z_k \partial z_l \partial z_m} =& \frac{\partial^2 P_{ic}}{\partial z_l \partial z_m} a_{ik} \left[ c - \sum_{c'=1}^{m_i} c' P_{ic'} \right] - \frac{\partial P_{ic}}{\partial z_l} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_m} \\ & - \frac{\partial P_{ic}}{\partial z_m} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_l} - P_{ic} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_l \partial z_m} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^4 P_{ic}}{\partial z_k \partial z_l \partial z_m \partial z_n} =& \frac{\partial^3 P_{ic}}{\partial z_l \partial z_m \partial z_n} a_{ik} \left[ c - \sum_{c'=1}^{m_i} c' P_{ic'} \right] - \frac{\partial^2 P_{ic}}{\partial z_l \partial z_m} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_n} \\ & - \frac{\partial^2 P_{ic}}{\partial z_l \partial z_n} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_m} - \frac{\partial P_{ic}}{\partial z_l} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_m \partial z_n} \\ & - \frac{\partial^2 P_{ic}}{\partial z_m \partial z_n} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_l} - \frac{\partial P_{ic}}{\partial z_m} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_l \partial z_n} \\ & - \frac{\partial P_{ic}}{\partial z_n} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_l \partial z_m} - P_{ic} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial^3 P_{ic'}}{\partial z_l \partial z_m \partial z_n}. \end{aligned}$$

The derivatives of $P_{ic}$ with respect to $a_{ij} \in \boldsymbol{a}_i$ and $b_{iv} \in \boldsymbol{b}_i$ are

$$\frac{\partial P_{ic}}{\partial a_{ij}} = P_{ic} c z_j - P_{ic} z_j \sum_{c'=1}^{m_i} c' P_{ic'},$$

and

$$\frac{\partial P_{ic}}{\partial b_{iv}} = \mathbf{1}(c \geq v) P_{ic} - P_{ic} \sum_{c'=v}^{m_i} P_{ic'}.$$

Then, we have

$$\frac{\partial^2 P_{ic}}{\partial z_k \partial u} = \frac{\partial P_{ic}}{\partial u} a_{ik} \left[ c - \sum_{c'=1}^{m_i} c' P_{ic'} \right] + \mathbf{1}(u = a_{ik}) P_{ic} \left[ c - \sum_{c'=1}^{m_i} c' P_{ic'} \right] - P_{ic} a_{ik} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial u},$$

$$\begin{aligned}
\frac{\partial^3 P_{ic}}{\partial z_k \partial z_l \partial u} =& \mathbf{1}(u = a_{ik}) \left[ \frac{\partial P_{ic}}{\partial z_l} \left( c - \sum_{c'=1}^{m_i} c' P_{ic} \right) - P_{ic} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_l} \right] \\
&+ a_{ik} \left[ \frac{\partial^2 P_{ic}}{\partial z_l \partial u} \left( c - \sum_{c'=1}^{m_i} c' P_{ic'} \right) - \frac{\partial P_{ic}}{\partial z_l} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial u} \right. \\
&\left. - \frac{\partial P_{ic}}{\partial u} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_l} - P_{ic} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_l \partial u} \right],
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^4 P_{ic}}{\partial z_k \partial z_l \partial z_m \partial u} =& \mathbf{1}(u = a_{ik}) \left[ \frac{\partial^2 P_{ic}}{\partial z_l \partial z_m} \left( c - \sum_{c'=1}^{m_i} c' P_{ic'} \right) - \frac{\partial P_{ic}}{\partial z_l} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_m} \right. \\
&\left. - \frac{P_{ic}}{\partial z_m} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_l} - P_{ic} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_l \partial z_m} \right] \\
&+ a_{ik} \left[ \frac{\partial^3 P_{ic}}{\partial z_l \partial z_m \partial u} \left( c - \sum_{c'=1}^{m_i} c' P_{ic'} \right) - \frac{\partial^2 P_{ic}}{\partial z_l \partial z_m} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial u} \right. \\
&- \frac{\partial^2 P_{ic}}{\partial z_l \partial u} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_m} - \frac{\partial P_{ic}}{\partial z_l} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_m \partial u} \\
&- \frac{\partial^2 P_{ic}}{\partial z_m \partial u} \sum_{c'=1}^{m_i} c' \frac{\partial P_{ic'}}{\partial z_l} - \frac{\partial P_{ic}}{\partial z_m} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_l \partial u} \\
&\left. - \frac{\partial P_{ic}}{\partial u} \sum_{c'=1}^{m_i} c' \frac{\partial^2 P_{ic'}}{\partial z_l \partial z_m} - P_{ic} \sum_{c'=1}^{m_i} c' \frac{\partial^3 P_{ic'}}{\partial z_l \partial z_m \partial u} \right].
\end{aligned}$$

The derivatives with respect to $\beta_{id}$ are equal to the product of $w_d$ and the derivatives with respect to $b_i$.

## Continuous data

The derivatives of $h_i$ with respect to $\boldsymbol{z}$ for continuous data (Huber et al., 2004) are given as follows. Note that only the first and second derivatives exist in this case.

$$\frac{\partial h_i}{\partial z_j} = \frac{a_{ij}}{\phi_i}(b_i + \boldsymbol{\beta}_i' \boldsymbol{w} + \boldsymbol{a}_i' \boldsymbol{z} - y_{if})$$

$$\frac{\partial^2 h_i}{\partial z_j \partial z_k} = \frac{a_{ij} a_{ik}}{\phi_i}$$

The derivatives of $h_i$ with respect to $a_{ij}$, $b_i$, and $\phi_i$ with continuous data are

$$\frac{\partial h_i}{\partial a_{ij}} = \frac{z_j}{\phi_i}(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z} - y_{if}),$$

$$\frac{\partial h_i}{\partial b_i} = \frac{(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z} - y_{if})}{\phi_i},$$

$$\frac{\partial h_i}{\partial \phi_i} = \frac{2y_{if}(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z}) - (b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z})^2 - t_i^2 + \phi_i}{2\phi_i^2},$$

$$\frac{\partial^2 h_i}{\partial z_j \partial a_{ix}} = \frac{a_{ij} z_x}{\phi_i} + \mathbf{1}(x = j)\frac{b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z} - y_{if}}{\phi_i},$$

$$\frac{\partial^2 h_i}{\partial z_j \partial b_i} = \frac{a_{ij}}{\phi_i},$$

$$\frac{\partial^2 h_i}{\partial z_j \partial \phi_i} = -\frac{a_{ij}(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z} - y_{if})}{\phi_i^2},$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial a_{ix}} = \frac{a_{ik}}{\phi_i}$$

and

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial \phi_i} = -\frac{a_{ij} a_{ik}}{\phi_i^2}.$$

The derivatives with respect to $\beta_{id}$ are equal to the product of $w_d$ and the derivatives with respect to $b_i$.

### Count data: Poisson distribution

The first- to fifth-order derivatives of $h_i$ with respect to $\boldsymbol{z}$ are

$$\frac{\partial h_i}{\partial z_j} = (\lambda_i - y_{if})a_{ij},$$

$$\frac{\partial^2 h_i}{\partial z_j \partial z_k} = a_{ij}\frac{\partial \lambda_i}{\partial z_k} = \lambda_i a_{ij} a_{ik},$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial z_l} = a_{ij}a_{ik}\frac{\partial \lambda_i}{\partial z_l} = \lambda_i a_{ij}a_{ik}a_{il},$$

$$\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m} = \lambda_i a_{ij}a_{ik}a_{il}a_{im}$$

and

$$\frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial z_n} = \lambda_i a_{ij}a_{ik}a_{il}a_{im}a_{in}.$$

The derivatives of $h_i$ with respect to $u \in \{\boldsymbol{a}_i, b_i\}$ are

$$\frac{\partial h_i}{\partial u} = (1 - \frac{y_{if}}{\lambda_i})\frac{\partial \lambda_i}{\partial u} = (\lambda_i - y_{if})z_j^{\mathbf{1}(u=a_{ij})},$$

$$\frac{\partial^2 h_i}{\partial z_j \partial a_{ix}} = a_{ij}\lambda_i z_x + (\lambda_i - y_{if})\mathbf{1}(x=j)$$

$$\frac{\partial^2 h_i}{\partial z_j \partial b_i} = a_{ij}\lambda_i,$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial a_{ix}} = a_{ij}a_{ik}\lambda_i z_x + \lambda_i a_{ik}\mathbf{1}(x=j) + \lambda_i a_{ij}\mathbf{1}(x=k),$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial b_i} = a_{ij}a_{ik}\lambda_i,$$

$$\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial a_{ix}} = a_{ij}a_{ik}a_{il}\lambda_i z_x + \lambda_i a_{ik}a_{il}\mathbf{1}(x=j) + \lambda_i a_{ij}a_{il}\mathbf{1}(x=k) + \lambda_i a_{ij}a_{ik}\mathbf{1}(x=l),$$

$$\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial b_i} = a_{ij}a_{ik}a_{il}\lambda_i,$$

$$\begin{aligned}\frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial a_{ix}} = {} & a_{ij}a_{ik}a_{il}a_{im}\lambda_i z_x + \lambda_i a_{ik}a_{il}a_{im}\mathbf{1}(x=j) + \lambda_i a_{ij}a_{il}a_{im}\mathbf{1}(x=k) + \\ & \lambda_i a_{ij}a_{ik}a_{im}\mathbf{1}(x=l) + \lambda_i a_{ij}a_{ik}a_{il}\mathbf{1}(x=m),\end{aligned}$$

and

$$\frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial b_i} = a_{ij}a_{ik}a_{il}a_{im}\lambda_i.$$

The derivatives with respect to $\beta_{id}$ are equal to the product of $w_d$ and the derivatives with respect to $b_i$.

**Count data: Negative-binomial distribution**

With a negative-binomial distribution, we have that

$$\begin{aligned} h_i = &- \log\Gamma\left(y_{if} + \frac{1}{\phi_i}\right) + \log(y_{if}!) + \log\Gamma\left(\frac{1}{\phi_i}\right) - y_{if}(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z}) \\ &+ y_{if}\log\left[\frac{1}{\phi_i} + \exp(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z})\right] + \frac{1}{\phi_i}\log\left[1 + \phi_i\exp(b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z})\right]. \end{aligned}$$

Let $\eta_i = b_i + \boldsymbol{\beta}_i'\boldsymbol{w} + \boldsymbol{a}_i'\boldsymbol{z}$. The derivatives with respect to $\boldsymbol{z}$ are

$$\frac{\partial h_i}{\partial z_j} = -y_{if}a_{ij} + (\phi_i y_{if} + 1)\frac{\exp(\eta_i)}{1 + \phi_i\exp(\eta_i)}a_{ij},$$

$$\frac{\partial^2 h_i}{\partial z_j \partial z_k} = (\phi_i y_{if} + 1)\frac{\exp(\eta_i)}{[1 + \phi_i\exp(\eta_i)]^2}a_{ij}a_{ik},$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial z_l} = -\,(\phi_i y_{if} + 1)\frac{\exp(\eta_i)[\phi_i\exp(\eta_i) - 1]}{[1 + \phi_i\exp(\eta_i)]^3}a_{ij}a_{ik}a_{il},$$

$$\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m} = (\phi_i y_{if} + 1)\frac{\exp(\eta_i)(\phi_i^2\exp(2\eta_i) - 4\phi_i\exp(\eta_i) + 1)}{[1 + \phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}a_{im},$$

and

$$\begin{aligned} \frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial z_n} = &-\,(\phi_i y_{if} + 1)\frac{\exp(\eta_i)\left[\phi_i^3\exp(3\eta_i) - 11\phi_i^2\exp(2\eta_i) + 11\phi_i\exp(\eta_i) - 1\right]}{[1 + \phi_i\exp(\eta_i)]^5} \\ &\times a_{ij}a_{ik}a_{il}a_{im}a_{in}. \end{aligned}$$

The derivatives with respect to $b_i$ are

$$\frac{\partial h_i}{\partial b_i} = -y_{if} + \left(y_{if} + \frac{1}{\phi_i}\right)\frac{\phi_i\exp(\eta_i)}{1 + \phi_i\exp(\eta_i)},$$

$$\frac{\partial^2 h_i}{\partial z_j \partial b_i} = (\phi_i y_{if} + 1)\frac{\exp(\eta_i)}{1 + \phi_i\exp(\eta_i)}a_{ij} - (\phi_i y_{if} + 1)\frac{\phi_i\exp(2\eta_i)}{[]1 + \phi_i\exp(\eta_i)]^2}a_{ij},$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial b_i} = (\phi_i y_{if} + 1)\frac{\exp(\eta_i)}{[1 + \phi_i\exp(\eta_i)]^2}a_{ij}a_{ik} - 2(\phi_i y_{if} + 1)\frac{\phi_i\exp(2\eta_i)}{[1 + \phi_i\exp(\eta_i)]^3}a_{ij}a_{ik},$$

$$\begin{aligned}\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial b_i} =& -(\phi_i y_{if}+1)\frac{\exp(\eta_i)[\phi_i\exp(\eta_i)-1]}{[1+\phi_i\exp(\eta_i)]^3}a_{ij}a_{ik}a_{il} - (\phi_i y_{if}+1)\frac{\phi_i\exp(2\eta_i)}{[1+\phi_i\exp(\eta_i)]^3}a_{ij}a_{ik}a_{il}\\ &+ 3(\phi_i y_{if}+1)\frac{\phi_i\exp(2\eta_i)[\phi_i\exp(\eta_i)-1]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial b_i} =&(\phi_i y_{if}+1)\frac{\exp(\eta_i)[\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)+1]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}a_{im}\\ &+ (\phi_i y_{if}+1)\frac{\exp(\eta_i)[2\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}a_{im}\\ &- 4(\phi_i y_{if}+1)\frac{\phi_i\exp(2\eta_i)[\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)+1]}{[1+\phi_i\exp(\eta_i)]^5}a_{ij}a_{ik}a_{il}a_{im}.\end{aligned}$$

The derivatives with respect to $a_{ic}$ are

$$\frac{\partial h_i}{\partial a_{ic}} = -y_{if}z_c + \left(y_{if}+\frac{1}{\phi_i}\right)\frac{\phi_i\exp(\eta_i)z_c}{1+\phi_i\exp(\eta_i)},$$

$$\begin{aligned}\frac{\partial^2 h_i}{\partial z_j \partial a_{ic}} =& -y_{if}\mathbf{1}(c=j) + (\phi_i y_{if}+1)\frac{\exp(\eta_i)}{1+\phi_i\exp(\eta_i)}a_{ij}z_c\\ &- (\phi_i y_{if}+1)\frac{\phi_i\exp(2\eta_i)}{[1+\phi_i\exp(\eta_i)]^2}a_{ij}z_c + (\phi_i y_{if}+1)\frac{\exp(\eta_i)}{1+\phi_i\exp(\eta_i)}\frac{a_{ij}}{a_{ic}}\mathbf{1}(c=j),\end{aligned}$$

$$\begin{aligned}\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial a_{ic}} =&(\phi_i y_{if}+1)\frac{\exp(\eta_i)}{[1+\phi_i\exp(\eta_i)]^2}a_{ij}a_{ik}z_c - 2(\phi_i y_{if}+1)\frac{\phi_i\exp(2\eta_i)}{[1+\phi_i\exp(\eta_i)]^3}a_{ij}a_{ik}z_c\\ &+ (\phi_i y_{if}+1)\frac{\exp(\eta_i)}{[1+\phi_i\exp(\eta_i)]^2}\frac{a_{ij}a_{ik}}{a_{ic}}\mathbf{1}(c\in\{j,k\}),\end{aligned}$$

$$\begin{aligned}\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial a_{ic}} =& -(\phi_i y_{if}+1)\frac{\exp(\eta_i)[\phi_i\exp(\eta_i)-1]}{[1+\phi_i\exp(\eta_i)]^3}a_{ij}a_{ik}a_{il}z_c\\ &- (\phi_i y_{if}+1)\frac{\phi_i\exp(2\eta_i)}{[1+\phi_i\exp(\eta_i)]^3}a_{ij}a_{ik}a_{il}z_c\\ &+ 3(\phi_i y_{if}+1)\frac{\phi_i\exp(2\eta_i)[\phi_i\exp(\eta_i)-1]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}z_c\\ &- (\phi_i y_{if}+1)\frac{\exp(\eta_i)[\phi_i\exp(\eta_i)-1]}{[1+\phi_i\exp(\eta_i)]^3}\frac{a_{ij}a_{ik}a_{il}}{a_{ic}}\mathbf{1}(c\in\{j,k,l\})\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial a_{ic}} =&(\phi_i y_{if}+1)\frac{\exp(\eta_i)[\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)+1]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}a_{im}z_c\\ &+(\phi_i y_{if}+1)\frac{\exp(\eta_i)[2\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}a_{im}z_c\\ &-4(\phi_i y_{if}+1)\frac{\phi_i\exp(2\eta_i)[\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)+1]}{[1+\phi_i\exp(\eta_i)]^5}a_{ij}a_{ik}a_{il}a_{im}z_c\\ &+(\phi_i y_{if}+1)\frac{\exp(\eta_i)[\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)+1]}{[1+\phi_i\exp(\eta_i)]^4}\frac{a_{ij}a_{ik}a_{il}a_{im}}{a_{ic}}\mathbf{1}(c\in\{j,k,l,m\}).\end{aligned}$$

The derivatives with respect to $\phi_i$ are, with $\psi$ denoting the digamma function,

$$\begin{aligned}\frac{\partial h_i}{\partial \phi_i} = -\frac{\log[1+\phi_i\exp(\eta_i)]}{\phi_i^2}+\left(y_{if}+\frac{1}{\phi_i}\right)\frac{\exp(\eta_i)}{1+\phi_i\exp(\eta_i)}-\frac{y_{if}}{\phi_i}+\\ \frac{\psi\left(y_{if}+\frac{1}{\phi_i}\right)}{\phi_i^2}-\frac{\psi\left(\frac{1}{\phi_i}\right)}{\phi_i^2},\end{aligned}$$

$$\frac{\partial^2 h_i}{\partial z_j \partial \phi_i} = y_{if}\frac{\exp(\eta_i)}{1+\phi_i\exp(\eta_i)}a_{ij}-(\phi_i y_{if}+1)\frac{\exp(2\eta_i)}{[1+\phi_i\exp(\eta_i)]^2}a_{ij},$$

$$\frac{\partial^3 h_i}{\partial z_j \partial z_k \partial \phi_i} = y_{if}\frac{\exp(\eta_i)}{[1+\phi_i\exp(\eta_i)]^2}a_{ij}a_{ik}-2(\phi_i y_{if}+1)\frac{\exp(2\eta_i)}{[1+\phi_i\exp(\eta_i)]^3}a_{ij}a_{ik},$$

$$\begin{aligned}\frac{\partial^4 h_i}{\partial z_j \partial z_k \partial z_l \partial \phi_i} =&-y_{if}\frac{\exp(\eta_i)[\phi_i\exp(\eta_i)-1]}{[1+\phi_i\exp(\eta_i)]^3}a_{ij}a_{ik}a_{il}-(\phi_i y_{if}+1)\frac{\exp(2\eta_i)}{[1+\phi_i\exp(\eta_i)]^3}a_{ij}a_{ik}a_{il}\\ &+3(\phi_i y_{if}+1)\frac{\exp(2\eta_i)[\phi_i\exp(\eta_i)-1]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^5 h_i}{\partial z_j \partial z_k \partial z_l \partial z_m \partial \phi_i} =&y_{if}\frac{\exp(\eta_i)[\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)+1]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}a_{im}\\ &+(\phi_i y_{if}+1)\frac{\exp(\eta_i)[2\phi_i\exp(2\eta_i)-4\exp(\eta_i)]}{[1+\phi_i\exp(\eta_i)]^4}a_{ij}a_{ik}a_{il}a_{im}\\ &-4(\phi_i y_{if}+1)\frac{\exp(2\eta_i)[\phi_i^2\exp(2\eta_i)-4\phi_i\exp(\eta_i)+1]}{[1+\phi_i\exp(\eta_i)]^5}a_{ij}a_{ik}a_{il}a_{im}.\end{aligned}$$

The derivatives with respect to $\beta_{id}$ are equal to the product of $w_d$ and the derivatives with respect to $b_i$.

# Appendices

# Appendix A

# Errata

Table A.1: Errata to Part I.

| Page | Line | Original Text | Correction Type | Corrected Text |
|---|---|---|---|---|
| 10 | 20 | 2.1.4.1 Gelstalt psychology perspective | Proofreading | 2.1.4.1 Gestalt psychology perspective |
| 10 | 21 | Gelstat psychologists emphasize... | Proofreading | Gestalt psychologists emphasize... |
| 11 | 34 | ...also knownas system identification... | Proofreading | ...also known as system identification... |
| 16 | 34 | Such information is used in Article I uses to... | Proofreading | Such information is used in Article I to... |
| 25 | 10 | ...Vista et al. (2016), and Vista et al. (2017), Zhu et al. (2016) viewed... | Proofreading | ...Vista et al. (2016), Vista et al. (2017), and Zhu et al. (2016) viewed... |
| 28 | 34 | ...into main two categories... | Proofreading | ...into two main categories... |
| 34 | 11 | ...maximize the marginal maximum likelihood function... | Proofreading | ...maximize the marginal likelihood function... |
| 59 | 33 | ...as GHQ and AGHA... | Proofreading | ...as GHQ and AGHQ... |
| 63 | 37 | ...process data usually has high dimensions. | Proofreading | ...process data usually have high dimensions. |