Method Versatility in Analysing Human Attitudes towards Technology

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktorin der Naturwissenschaften
(Dr. rer. nat.)
genehmigte Dissertation

von Frau
Olga Lezhnina, M.Sc.
geboren am
05.07.1971
in
Krasnodar, Russland

2023

1. Referent: Prof. Dr. Sören Auer
2. Referent: Prof. Dr. Stefan Mol

Tag der Promotion: 06.07.2023

Abstract

Various research domains are facing new challenges brought about by growing volumes of data. To make optimal use of them, and to increase the reproducibility of research findings, method versatility is required. Method versatility is the ability to flexibly apply widely varying data analytic methods depending on the study goal and the dataset characteristics.

Method versatility is an essential characteristic of data science, but in other areas of research, such as educational science or psychology, its importance is yet to be fully accepted. Versatile methods can enrich the repertoire of specialists who validate psychometric instruments, conduct data analysis of large-scale educational surveys, and communicate their findings to the academic community, which corresponds to three stages of the research cycle: measurement, research *per se*, and communication. In this thesis, studies related to these stages have a common theme of human attitudes towards technology, as this topic becomes vitally important in our age of ever-increasing digitization.

The thesis is based on four studies, in which method versatility is introduced in four different ways: the consecutive use of methods, the toolbox choice, the simultaneous use, and the range extension. In the first study, different methods of psychometric analysis are used consecutively to reassess psychometric properties of a recently developed scale measuring affinity for technology interaction. In the second, the random forest algorithm and hierarchical linear modeling, as tools from machine learning and statistical toolboxes, are applied to data analysis of a large-scale educational survey related to students' attitudes to information and communication technology. In the third, the challenge of selecting the number of clusters in model-based clustering is addressed by the simultaneous use of model fit, cluster separation, and the stability of partition criteria, so that generalizable separable clusters can be selected in the data related to teachers' attitudes towards technology. The fourth reports the development and evaluation of a scholarly knowledge graph-powered dashboard aimed at extending the range of scholarly communication means.

The findings of the thesis can be helpful for increasing method versatility in various research areas. They can also facilitate methodological advancement of academic training in data analysis and aid further development of scholarly communication in accordance with open science principles.

Keywords: method versatility, data science, attitudes towards technology.

Zusammenfassung

Verschiedene Forschungsbereiche müssen sich durch steigende Datenmengen neuen Herausforderungen stellen. Der Umgang damit erfordert – auch in Hinblick auf die Reproduzierbarkeit von Forschungsergebnissen – Methodenvielfalt. Methodenvielfalt ist die Fähigkeit umfangreiche Analysemethoden unter Berücksichtigung von angestrebten Studienzielen und gegebenen Eigenschaften der Datensätze flexible anzuwenden.

Methodenvielfalt ist ein essentieller Bestandteil der Datenwissenschaft, der aber in seinem Umfang in verschiedenen Forschungsbereichen wie z. B. den Bildungswissenschaften oder der Psychologie noch nicht erfasst wird. Methodenvielfalt erweitert die Fachkenntnisse von Wissenschaftlern, die psychometrische Instrumente validieren, Datenanalysen von groß angelegten Umfragen im Bildungsbereich durchführen und ihre Ergebnisse im akademischen Kontext präsentieren. Das entspricht den drei Phasen eines Forschungszyklus: Messung, Forschung *per se* und Kommunikation. In dieser Doktorarbeit werden Studien, die sich auf diese Phasen konzentrieren, durch das gemeinsame Thema der Einstellung zu Technologien verbunden. Dieses Thema ist im Zeitalter zunehmender Digitalisierung von entscheidender Bedeutung.

Die Doktorarbeit basiert auf vier Studien, die Methodenvielfalt auf vier verschiedenen Arten vorstellt: die konsekutive Anwendung von Methoden, die Toolbox-Auswahl, die simultane Anwendung von Methoden sowie die Erweiterung der Bandbreite. In der ersten Studie werden verschiedene psychometrische Analysemethoden konsekutiv angewandt, um die psychometrischen Eigenschaften einer entwickelten Skala zur Messung der Affinität von Interaktion mit Technologien zu überprüfen. In der zweiten Studie werden der Random-Forest-Algorithmus und die hierarchische lineare Modellierung als Methoden des Machine Learnings und der Statistik zur Datenanalyse einer groß angelegten Umfrage über die Einstellung von Schülern zur Informations- und Kommunikationstechnologie herangezogen. In der dritten Studie wird die Auswahl der Anzahl von Clustern im modellbasierten Clustering bei gleichzeitiger Verwendung von Kriterien für die Modellanpassung, der Clustertrennung und der Stabilität beleuchtet, so dass generalisierbare trennbare Cluster in den Daten zu den Einstellungen von Lehrern zu Technologien ausgewählt werden können. Die vierte Studie berichtet über die Entwicklung und Evaluierung eines wissenschaftlichen wissensgraphbasierten Dashboards, das die Bandbreite wissenschaftlicher Kommunikationsmittel erweitert.

Die Ergebnisse der Doktorarbeit tragen dazu bei, die Anwendung von vielfältigen Methoden in verschiedenen Forschungsbereichen zu erhöhen. Außerdem fördern sie die methodische Ausbildung in der Datenanalyse und unterstützen die Weiterentwicklung der wissenschaftlichen Kommunikation im Rahmen von Open Science.

Schlagwörter: Methodenvielfalt, Datenwissenschaft, Einstellung zu Technologien.

Acknowledgements

I would like to express my gratitude to everyone who supported me in conducting these studies and writing the thesis. First and foremost, this work would not have been possible without the assistance of my dear family, people who always inspire me with their example, bring meaning and delight in everything we do together, and support me in my scientific work and in any vicissitudes of life.

The supervisors of this thesis, Prof Dr Sören Auer and Dr Gábor Kismihók, provided me with everything that a researcher and a PhD candidate could ever wish for: unstinting encouragement and realistic feedback, eagerness to cooperate and respect for my independent working style. They were always there so that I could ask for advice, discuss an idea, or get a spark of inspiration from our talks about data analysis, scholarly knowledge graphs, or any other topic. For me, Prof Dr Sören Auer is a stimulating example of an extraordinary knowledgeable scientist, the impression which deepened after I attended a course of his lectures at the LUH and admired his patience, empathy, and dedication. Exhilarating conversations with Dr Gábor Kismihók let me look at familiar phenomena from an unusual perspective, immerse in yet unstudied areas of scientific methodology, and brighten my spirits. Organisational support, which both supervisors never failed to offer, gave me a possibility to feel safe and focus at research.

Thought-stimulating discussions with my colleagues at the TIB at research seminars and informal meetings were useful for many aspects of this work, such as conceptualising my approach and formulating the methodological strategies. Everyone was willing to help when asked for any kind of assistance, or discuss a topic they are proficient at. Such encounters most frequently happened within our team at Learning and Skill Analytics group, and the group members created a very amiable atmosphere at the workplace. In particular, I would like to express my sincere gratitude to Dr Brian Cahill for looking at this text from a native speaker perspective and making useful suggestions. A very special role in improving the four manuscripts, and the thesis in general, was played by Manuel Prinz, who provided me with immensely helpful feedback on each of the papers, co-authored the fourth publication, and commented on the thesis.

Outside the TIB, there are a few professionals who I specifically wish to express my gratitude to. Prof Dr Moritz Heene from the LMU, Munich, impressed me with his devotion to open science practices and ability to convince his students and colleagues that statistics is fun. Also in Munich, Dr Elena Gaertner supported me in research and professional choices, and I am deeply indebted to her. Sylvia Pittroff, who can discuss a wide spectrum of ideas, or detect any tiny discrepancy in the text and point it out most nicely, has helped my scientific endeavour greatly. Dr Daniel Wessel, as I mention in Chapter 4, generously shared the data that at that time was not published yet. Most reviews in the journals to which I submitted the papers were useful, and I appreciate the dedication of the anonymous reviewers. Many thanks to everyone who supported this work with their time, energy, and intellectual effort.

Contents

Chapter 1

Introduction

Versatility as a wide diversity of capabilities was discussed in relation to scientific thinking as early as in 1880s [241]. Since then, philosophers of science have returned to this topic by emphasizing the importance of a pluralistic approach to scientific methodology [51], and statisticians by discussing multiple competing perspectives on statistical inference and decision making [84]. In this thesis, I explore the epistemological problem of versatility in relation to data analytic methods. The term "method" is used broadly to describe general strategies of data analysis, as well as specific models, algorithms, tests, and metrics. Thus, method versatility is understood as the ability to flexibly apply widely varying data analytic methods depending on the study goal and the dataset characteristics.

Method versatility can be viewed as one of the pivotal concepts in scientific research. In their milestone paper on methodological problems in statistical analysis, Gelman and Hennig [85] suggested a framework of scientific principles. In Figure 1.1, I grouped them to depict their relevance to the aims of this thesis.

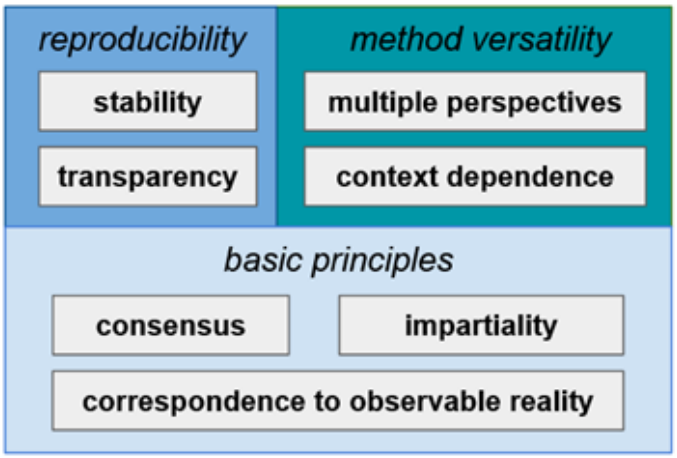

Figure 1.1: **Method Versatility in the Framework of Scientific Principles.**

The basic principles include a consensus about definitions and assumptions; impartiality of researchers; and correspondence of theories to observable reality that can confirm theoretical premises or refute them. Without these most fundamental principles, scientific research is not possible, and this work does not focus on them. Stability of scientific findings contributes to the reproducibility of research results, and transparent reporting is vital for methodological rigour [259]. Therefore, I adhere to maximal transparency in reporting my methods and results. Finally, awareness of multiple perspectives and context dependence are principles directly related to method versatility. These principles mean recognizing that various perspectives on analysis are a reality to be reckoned with, and the choice of a suitable method depends on the aims of the study and on the selected datasets. Thus, method versatility, without being a panacea against all possible flaws in data analysis, is a key ingredient of methodologically rigorous and scientifically valid research.

The topic of method versatility becomes increasingly important in our age of big data. Most inferential statistical tests and machine learning (ML) models have a requirement of a minimal sample size [71], and the choice of methods for studies with small sample sizes is inevitably limited. The situation has changed with the growing amount of data that can be analysed with various methods to obtain new insights. As early as in 2001, Leo Breiman said: “To solve a wider range of data problems, a larger set of tools is needed” [29]. Since then, the evolving field of data science has brought forth advanced instruments for different tasks. From the early days of data science, method versatility has been its essential characteristic [35], and it is a common understanding that a suitable method should be selected for each specific task [1]. In other areas of research, however, such as social sciences, psychology, and educational science, the importance of method versatility is yet to be fully accepted. In these domains, researchers frequently resort to a rather restricted range of instruments, or even a single “default method” [84], which they are most acquainted with due to their training and practice. The insufficiently flexible approach to data analysis in these areas was discussed as a problem to be resolved in order to deal with the replication crisis [47], and a context-based “toolbox approach” was called for [86]. Social and educational sciences have begun to utilize data science perspectives, and new subfields have emerged that practice versatile methods of data analysis [211]. This endeavour can be reciprocal, as it was shown that threats to generalizability of findings are quite similar in diverse areas of research [121], and to deal with these challenges, interdisciplinary effort might be required.

1.1 Motivation, Problem Statement, and Challenges

Data science has evolved as an interdisciplinary field that synthesizes statistics, informatics, computer science, and communication, with the aim to study data and obtain insights that can be transferred to the related areas. In its further development, it requires dynamic exchange with other areas in accordance with the principles of open science [59]. Therefore, it is vital for data science and for other scientific domains to promote cross-domain activities and develop a synergy of research disciplines. This

approach creates opportunities for theoretical and technological advancements that can address scientific problems more effectively than can be achieved by singular disciplinary efforts [35].

To provide domain-specific research with data science methods and strategies, it might be useful to differentiate stages of the research cycle, which can be conceptualized in various ways (see, e.g., [96], [131]). A simple tripartite cycle is presented in Figure 1.2.

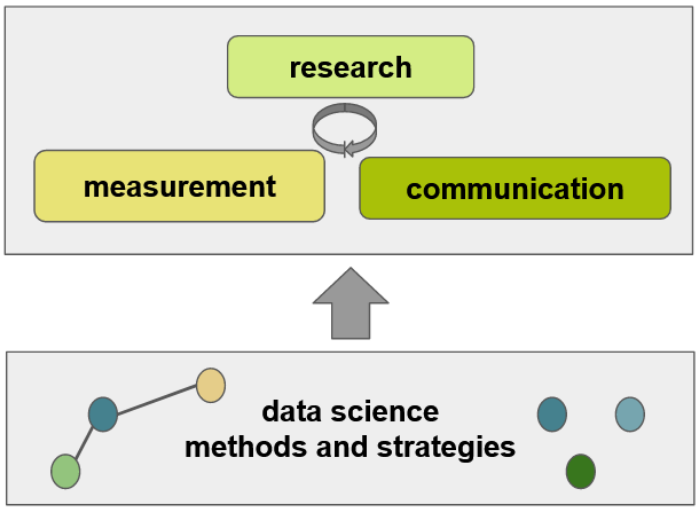

Figure 1.2: **The Domain Research Cycle and Data Science Input.**

Measurement is necessary to gather data, and appropriate instruments are required for this stage. In the natural sciences, physical or chemical properties of research objects are measured with technical instruments of ever-increasing precision. In social sciences, psychology, educational science, and certain subfields of medicine, which deal with subjective human experience, psychometric instruments can be used. These are scales aimed at capturing a psychological construct, usually with a few items, although one-item scales also exist. The construct under scrutiny is understood as latent, that is, it cannot be assessed directly [261], and therefore, perspectives on measurement precision and ways to increase it might differ. Psychometric assessment of a scale is conducted to establish its usefulness as a measurement instrument, and this process requires data analysis with various methods.

The next stage is conducting research *per se*, which is a general category for scientific work different from psychometric analysis and scholarly communication. This stage consists of either acquiring new insights from the data or testing previously formulated hypotheses [229]. Such studies are conducted in all scientific disciplines, and the challenges that researchers face depend to some degree on their domain. In this work, I narrow the area of my attention and refer mostly to domain-specific research in educational science, specifically, to data analysis of large-scale educational surveys.

These surveys gather information on various teachers' and students' characteristics, including aptitude, skills, demographics, and psychological attributes [76], [181]. They provide researchers with the large amount of high quality data that can be analysed with various statistical and ML procedures and therefore are suitable for research on method versatility.

Finally, researchers need to share their findings with the academic community. Recent advancement in open science practices, such as guidelines for transparent reporting, are aimed at improving scholarly communication. Works on novel methods of information retrieval, such as knowledge graphs (KGs), also belong to this field. Scholarly knowledge graphs (SKGs) increase accessibility and machine readability of research findings [13] and support scholarly communication in accordance with FAIR (findable, accessible, interoperable, reusable) principles [96]. For wider acceptance of this novel technology, user-friendly interfaces need to be developed, and in this work, I focus on SKG-based interfaces as an aid to scholarly communication. This stage of the research cycle can also benefit from method versatility that might be facilitated with the assistance of data science; here, in contrast to other stages, extending the range of available methods could be related to novel ways of presenting academic findings rather than data analysis.

In addition to conceptualisation of the research cycle, a common thread integrating these three stages would be helpful for structuring the work, and for this purpose, I selected the topic of human attitudes towards technology. An attitude is a multifaceted structure comprised of affective, cognitive, and behavioural components [97]; human attitudes towards technology include a wide spectrum of responses, from aversion and anxiety to interest and enjoyment. This topic is studied by information science, computer science, educational science, and many other domains, as human attitudes towards technology are vital for learning, professional activities, and personal wellbeing in our digitalized world [77]. At the measurement stage, psychometric instruments related to a specific attitude towards technology (such as affinity or interest), or a set of positive and negative attitudes, can be developed and validated. The research stage gives a possibility to explore relationships between respondents' attitudes towards technology and other characteristics (such as students' academic achievements in large-scale educational surveys) and acquire new insights from the data containing these variables [11], [153], [191]. The role of human attitudes towards technology at the communication stage is twofold. It can be a topic of communication, but these attitudes should be also taken into account when novel means of scholarly communication are developed, as knowledge of technology acceptance principles helps making them appealing for users [115].

In this thesis, I explore possible ways of introducing a wider range of flexibly used methods in analysing human attitudes towards technology at the stages of measurement, research, and communication. My motivation in conducting this work is to find versatile ways of facilitating method versatility and bring data science perspectives into such domains as psychometric analysis, educational data analysis, and

scholarly communication. This individual input, however modest, might be an addition to cumulative effort of researchers working towards increasingly rigorous scientific methods. The research problem of this thesis can be thus formulated as follows:

Research problem definition:

How can method versatility in data analysis of human attitudes towards technology be facilitated at each stage of the research cycle?

1.1.1 Research Challenges in Method Versatility

For addressing this problem, it is necessary to overview the current situation in relevant domains. The following research gaps, or challenges faced by researchers in these areas, were identified to be considered in conducting this work:

Challenge 1. Underuse of diverse methods in psychometric analysis related to human attitudes towards technology

To measure human attitudes towards technology, valid and reliable psychometric instruments are required. Theoretically, a scale should undergo thorough evaluation, not only in the process of development, but repeatedly afterwards, with different populations, to assess its psychometric properties, such as validity, dimensionality, reliability, and item functioning. Unfortunately, a tendency persists to use instruments that lack this important information: many scales still have unclear validity [16], as the results of their evaluation were not reported transparently [113]. Such a situation contributes to difficulties with the replicability of research [72]. One of the causes of these problems is insufficient variety of methods and adherence to a rather limited range of standard procedures, most often to methods of Classical Test Theory (CTT). These default methods typically used in psychometric analysis often give biased results due to various reasons, such as neglected assumption checking [93]. Better practices in scale validation are required to overcome the current situation in the field, which was called the validation crisis [220]. Using diverse methods, including those developed in the frame of parametric and nonparametric Item Response Theory (IRT), is beneficial for this task [58].

Challenge 2. Insufficient methodological flexibility in educational data analysis related to human attitudes towards technology

Data analysis of large-scale educational surveys can be a challenging task for researchers in relevant domains [8]. For this purpose, supervised and unsupervised learning methods are used: the former deals with labelled data and includes classification and regression, and the latter with unlabelled data (for instance, cluster analysis). In educational data analysis, the use of an insufficiently wide range of methods has been discussed as a problem, which often leads to oversimplified or biased

interpretations [271]. It is less frequently the case when a pipeline of analysis for a specific large-scale educational survey is formalized (see, e.g., [178]), but many analytical decisions are to be made by researchers themselves, and some commonly used default methods might be ineffective. In particular, the hierarchical structure of the data should be taken into account, as the indiscriminate use of non-multilevel default methods leads to an increase in the Type I error rate [44]. Dealing with missing data is also a typical problem for data analysis of large-scale educational surveys, which should be approached with rigour and flexibility [91].

Challenge 3. Restricted range of means of scholarly communication related to human attitudes towards technology

Shortcomings of contemporary scholarly communication, which lead to the replication crisis, include limited findability of research; unequal access to published papers; deterioration of peer review quality; compromised research integrity; insufficient machine readability of literature; and restricted availability of open research tools [96]. These problems can be to a substantial degree resolved by SKG, which is a novel modality of information retrieval. Wider use of SKGs in academia is beneficial for the digitalization of published works and further development of scholarly communication. However, researchers in many disciplines still do not engage to a sufficient degree in the use of this innovative method of communication [13], and making SKGs more appealing to a wider academic community is a goal that has yet to be attained [215].

In specifying the challenges of this work, I narrowed down the areas in which method versatility can be facilitated. The results are presented in Figure 1.3.

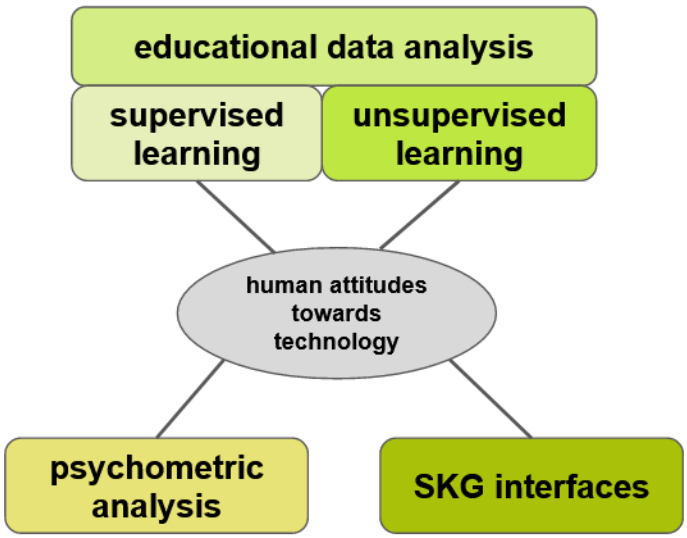

Figure 1.3: **Four Areas for Facilitating Method Versatility.**

My task thus becomes more specific: to explore the ways of facilitating method versatility in validation of psychometric instruments; in supervised and unsupervised learning related to data analysis of large-scale educational surveys; and in development of SKG-based interfaces, with studies in each of these areas focusing on human attitudes towards technology.

1.1.2 Approach

In order to facilitate method versatility, an approach should be found that can be reproduced by other researchers. My approach is to suggest strategies for widening the spectrum of commonly applied methods in order to make research findings more generalizable and reproducible.

This approach is different from one typically used in social sciences and aligned domains, when a researcher develops a scale, uses it to conduct a study, and presents the findings to the public. Although I also deal with each of the stages of the research cycle, my focus is on methods rather than findings; the results such as a validated scale, or an established relationship between variables, which are most interesting for domain experts, are perceived as secondary in the methodological research that I undertake. My approach also differs from those typically used to compare the effectiveness of various methods, e.g., clustering algorithms, or to introduce a novel method and assess its performance against the baseline. Although I show the benefits of suggested strategies in comparison to commonly used default procedures, I resort to already existing methods that I select based on already existing findings of their effectiveness. The approach is similar to what was either called for, with conceptual explanations, or explored as a practical solution to a specific problem in previous studies, which are discussed in detail in section 2.1 and section 3.1 of this thesis.

Thus, I suggest an integration of already existing methods into context-dependent adjustable analytical strategies. In that, I adhere to a set of principles depicted in Figure 1.4. These can be summarized as follows:

(1) Usefulness: a strategy ought to address a research gap, or a methodological problem, and be of use to domain specialists, data analysts, and, hopefully, a wider academic community;
(2) Ease-of-use: a strategy needs to be easily reproducible by researchers with different levels of proficiency in data analysis;
(3) Variability: strategies in different contributions should introduce method versatility in different ways, so that the audience is exposed to various means towards the goal;
(4) Modularity: a strategy ought to consist of blocks and levels, so that each of them can be used separately, or all together;
(5) Transparency: the strategies and their implementations need to be reported with maximal transparency for better reproducibility.

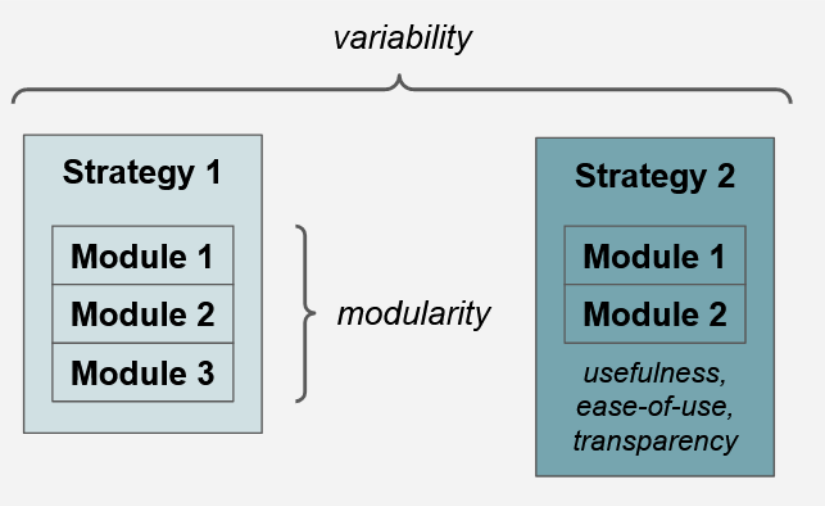

Figure 1.4: **Principles of the Approach Applied in the Thesis.**

1.2 Research Questions

The approach outlined above is applied to research questions addressed in this thesis. Based on the challenges faced by relevant fields, the following research questions were formulated:

RQ1. How to facilitate method versatility in validation of psychometric instruments related to human attitudes towards technology?

In order to proceed with RQ1, existing ways of dealing with the problems in the area should be explored. In previous studies on psychometric analysis, methods aimed at overcoming limitations of commonly used practices were developed (see, e.g., [36], [238], [261]). An analytical approach was suggested that combines CTT and IRT techniques in assessment of dimensionality, reliability, and item functioning of a scale [58]. In my dealing with the RQ1, I use this approach and make amendments to it. As it does not include tests of validity, finding a way to assess construct validity of a scale by means other than the commonly used default options [22] is required. For practical application of the strategies, I select a recently developed psychometric instrument measuring human attitudes towards technology, which was validated by its authors with commonly used methods [80].

Thus, in relation to RQ1, I explore a possible way of facilitating method versatility in evaluation of validity, dimensionality, reliability, and item functioning of a scale measuring human attitudes towards technology.

RQ2. How to facilitate method versatility in educational research on human attitudes towards technology?

In regard to RQ2, I attempt to find an opportunity to facilitate method versatility in educational data analysis; specifically, in data analysis of large-scale educational surveys, such as the Programme for International Student Assessment (PISA) and the International Computer and Information Literacy Study (ICILS). In these surveys, the topic of human attitudes towards technology is presented by frameworks assessing respondents' attitudes towards information and communication technology (ICT), and sufficiently large sample sizes allow for variability in analytical decisions.

I deal with supervised (classification and regression) and unsupervised (clustering) learning tasks for educational data analysis. In case of supervised learning, I use methods from ML and statistical toolboxes that are most suitable for specific purposes. Flexible and context-based use of these two sets of methods was discussed in literature [29], [229] but yet to be attained in data analysis of large-scale educational surveys. In the area of cluster analysis, I explore the topic of selecting the number of clusters in model-based clustering. It is a challenging task that can be handled by integration of different perspectives [7], [111]. Researchers often rely excessively on fit indices, as model fit is the main selection criterion in model-based clustering; it was shown, however, that a wider spectrum of criteria needs to be taken into account. Thus, to address RQ2, I explore the ways of introducing method versatility for supervised and unsupervised learning tasks in data analysis of large-scale educational surveys in relation to human attitudes towards technology.

RQ3. How to facilitate method versatility in communication of research results related to human attitudes towards technology?

For the wider application of SKGs, it is necessary to make it more appealing to research communities in various academic areas [13], [215]. Therefore, I explore a possibility of developing an easy-to-use interface based on the theoretical premises of technology acceptance. Previous research stressed the importance of visual interfaces which employ principles of computer science, graphic design, and human-technology interaction [38]. Thus, to deal with RQ3, I attempt to facilitate method versatility in communication of research results related to human attitudes towards technology.

1.3 Thesis Overview

This section summarizes four contributions of the thesis. I aimed at facilitating method versatility at three stages of the research cycle: measurement, research *per se*, and communication of results. Human attitudes towards technology were the topic of

measurement, research, or communication in each of the four contributions. In the fourth contribution, which is related to RQ3 (the communication stage), these attitudes were also taken into account in the process of the interface development.

In conducting this work, I adhered to principles that were outlined as formative for my approach (section 1.2). I intended the suggested strategies for as wide audience as possible, and in particular, made my code publicly available on GitHub and wrote it in R, which is used by social and educational scientists more commonly than other programming languages. The only exception is the contribution to RQ3, in which I implemented a web service in Python using the Flask framework and JavaScript (JS); the code, however, is simple and reproducible. The links to repositories can be found in the respective chapters (sections 4.1, 5.1, 6.1, 7.1).

1.3.1 Contributions

This section presents four contributions to the thesis. The contributions and their relation to RQs are shown in Figure 1.5.

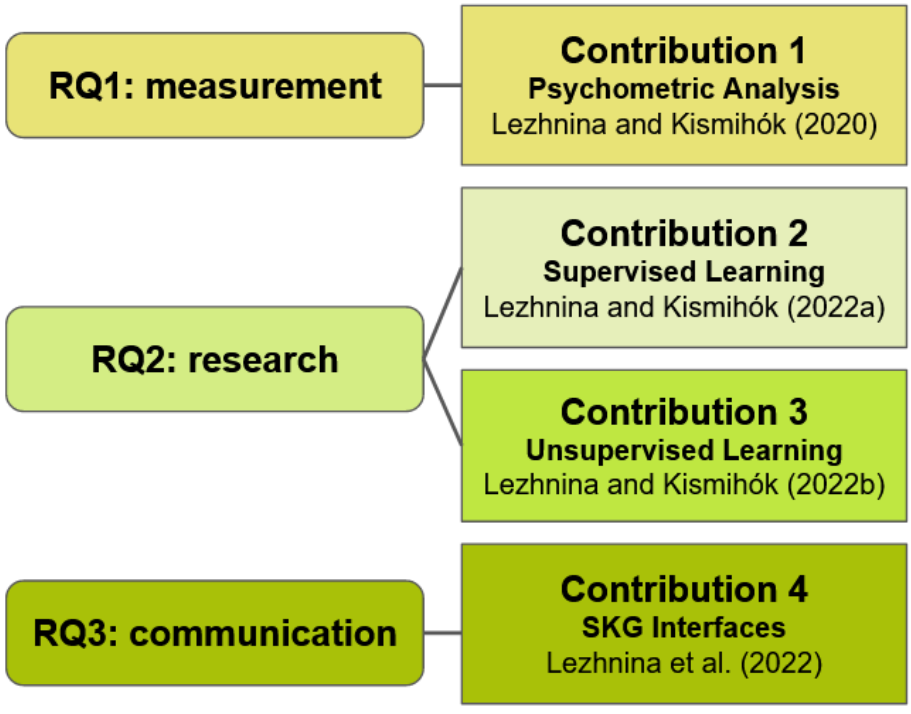

Figure 1.5: **Research Questions and Thesis Contributions.**

For each RQ, a study was conducted and published in an international peer reviewed journal; for the RQ2, two studies were carried out that cover supervised and unsupervised learning methods.

Contributions to RQ1

To address RQ1, I explored a possibility to reassess the results of a scale validation with different methods. To illustrate the strategy that I suggested, I selected a recently developed psychometric instrument measuring human attitudes towards technology, the Affinity for Technology Interaction (ATI) scale. The authors of the scale, who kindly shared their data for reassessment, validated it by means of CTT, with construct validity indicated by results of correlation analysis [80]. In my approach, methods of parametric and nonparametric IRT were applied for dimensionality, reliability, and items functioning as shown in [58] with amendments related to assumption checking. For validity analysis, I applied hierarchical clustering of variables as an alternative method presenting the results with various levels of granularity [45], selected an implementation of the clustering approach that is based on Principal Component Analysis (PCA) [43], and explored the stability of cluster partitions. Due to the suggested strategy, the findings on validity, dimensionality, reliability, and item functioning of the scale were reconfirmed; more detailed information about item-level characteristics was obtained and communicated to the authors of the scale; and versatile methods of psychometric analysis were suggested to researchers in the area.

Contributions to RQ2

In relation to RQ2, I conducted two studies dealing with (i) supervised learning (classification and regression tasks) and (ii) unsupervised learning (cluster analysis). Both studies focused on method versatility in educational data analysis, and in both, human attitudes towards technology were the topic of research.

In the first study, I applied ML and statistical methods to explore German students' attitudes towards ICT in relation to their academic achievements measured by PISA in 2015 and 2018. I used the random forest (RF) algorithm for missing data imputation and for predicting students' proficiency levels in mathematics and science (the classification task). Hierarchical linear modelling (HLM) was applied to explore associations between students' scores and their attitudes towards ICT (the regression task). The study provides researchers with detailed explanation of the strategy involving the flexible use of instruments from ML and statistical toolboxes that were most suitable for the specific tasks.

The second study focused on selecting the number of clusters in model-based clustering. I suggested an extended analytical strategy for selecting the number of clusters in Latent Class Cluster Analysis (LCCA) by integrating model-based and distance-based criteria with the bootstrap stability assessment. The suggested strategy of simultaneous use of these criteria was illustrated on the simulated data and on the real-world dataset from the ICILS 2018. I used the data of German teachers' attitudes

to ICT and showed that the extended strategy, in comparison to fit indices-based strategy, facilitates the selection of more stable and well-separated clusters in the data.

Contributions to RQ3

To deal with RQ3, I explored the ways of increasing versatility in communication of research results via SKG as novel modality of information retrieval. To increase the acceptance of SKGs and extend the range of SKG-based interfaces, I developed a dashboard, which visualizes research contributions on attitudes towards ICT in PISA 2015 and 2018 in the frame of the Open Research Knowledge Graph (ORKG) research service infrastructure initiative. According to preliminary results of the user evaluation survey, the dashboard was perceived as more appealing than the baseline ORKG-powered interface. These findings can be used for the development of SKG-powered dashboards in different domains, thus facilitating acceptance of these novel instruments by research communities and increasing versatility in scholarly communication.

1.3.2 Publications

The thesis is based on the following publications in peer reviewed international journals:

Lezhnina, O., & Kismihók, G. (2020). A multi-method psychometric assessment of the Affinity for Technology Interaction (ATI) Scale. *Computers in Human Behavior Reports, 1*, Article 100004. https://doi.org/10.1016/j.chbr.2020.100004

Lezhnina, O., & Kismihók, G. (2022a). Combining statistical and machine learning methods to explore German students' attitudes towards ICT in PISA. *International Journal of Research & Method in Education, 45*(2), 180–199. https://doi.org/10.1080/1743727X.2021.1963226

Lezhnina, O., & Kismihók, G. (2022b). Latent class cluster analysis: Selecting the number of clusters. *MethodsX, 9*, Article 101747. https://doi.org/10.1016/j.mex.2022.101747

Lezhnina, O., Kismihók, G., Prinz, M., Stocker, M., & Auer, S. (2022). A scholarly knowledge graph-powered dashboard: Implementation and user evaluation. *Frontiers in Research Metrics and Analytics, 7*, Article 934930. https://doi.org/10.3389/frma.2022.934930

Throughout this thesis, the pronoun "I" is used, which by no means implies underestimating the investments of my co-authors that are recognized in author contribution statements in each of the publications. The singular first-person pronoun is used to emphasize that as the author of conceptual approach, methodological implementation, and the code, I take responsibility for the findings and shortcomings of

this work. Appreciation for invaluable help of other parties is expressed in the Acknowledgements section of the thesis.

1.4 Thesis Structure

The thesis includes eight chapters. The first three chapters discuss the basic theoretical concepts and existing literature on the topic; in the following four chapters, I report the findings of studies I conducted in the frame of the thesis, and the last chapter summarises the results of this work.

In Chapter 1, I introduce the topic of the thesis, define the main concepts, such as method versatility, and explain the motivation for facilitating it at all stages of the research cycle: measurement, research, and communication. I formulate the research problem, the research questions, and the approach used in this work. After that, I briefly sketch four contributions to the thesis. Chapter 2 contains the theoretical and methodological background for the thesis. It starts with discussing method versatility and possible ways of applying it to data analysis. Theoretical premises of psychometric research, data analysis of large-scale educational surveys (supervised and unsupervised learning), and scholarly communication via SKGs are outlined. In the last section of the chapter, I define the topic of the four contributions, human attitudes towards technology, and discuss related frameworks used in this thesis, such as affinity for technology, ICT engagement in PISA, and views on ICT in ICILS. Chapter 3 describes previous work on which the thesis is based. In terms of method versatility, these are theoretical calls for multi-method data analysis in related areas and practical approaches that can be incorporated into the strategies which I suggest. In psychometric analysis, these are previous works on combining CTT and IRT and on hierarchical clustering for validity analysis. In educational data analysis, these are statistical and ML methods applied to large-scale educational surveys and cluster selection procedures in LCCA. In the area of SKG interfaces, these are previously developed graph-based dynamic visualisations.

In Chapter 4, I give the details of the first contribution, a psychometric study conducted to address RQ1. I report the consecutive use of CTT and IRT methods to assess dimensionality, reliability, and item functioning of the ATI scale, and hierarchical clustering of variables for validity analysis. In Chapter 5 and Chapter 6, I deal with RQ2 and discuss data analysis of large-scale educational surveys. In Chapter 5, I suggest using the toolbox choice of statistical and ML methods to analyse German students' attitudes to ICT in PISA 2015 and 2018. Implementations of the RF algorithm were applied to missing data imputation and the classification task, while for the regression task, a statistical method, HML, was used. In Chapter 6, I suggest an extended strategy for selecting the number of clusters in LCCA, which implies the simultaneous use of model fit, cluster separation, and stability of the partitions criteria. The strategy is illustrated on the simulated data and on the German subset of teachers' views on ICT from ICILS 2018. Chapter 7 addresses the problem of method versatility in scholarly communication related to RQ3 by extending the range of SKG-based

visualisations. I report developing and evaluating an ORKG-based dashboard that visualises research results on attitudes to ICT in PISA. Finally, Chapter 8 reiterates the research questions with information about impact of the conducted work on methodological development in related areas. In this final chapter, I address limitations of the thesis and indicate directions for further research.

Chapter 2

Background

In this chapter, I give an overview of the main theoretical notions underlying methods used in this thesis. I start by discussing method versatility and different ways of introducing it into data analysis. Then, I outline basic concepts used in the thesis. They are related to the areas of psychometric analysis, supervised and unsupervised learning methods applied to large-scale educational surveys, and scholarly communication via SKGs. Finally, I explain how human attitudes towards technology are defined, measured, and studied in contexts related to this work.

2.1 Method Versatility in Data Analysis

In data science, versatility is deeply ingrained into methodological approaches and pervades routine practices [35]. It is inherent to the data science perspective to flexibly choose methods for a specific task; as we know from "No Free Lunch" theorems for search, optimization, and supervised learning, no algorithm can outperform others on all types of problems [1]. It is not the case in many other disciplines, though. The problem might be to some degree related to a possible misconception by which flexible context-dependent use of methods thoroughly selected from a wider spectrum of options is confused with "undisclosed flexibility of analytical choices" [121], [230]. In social sciences, adherence to a restricted range of default methods has been recognized as a methodological shortcoming, and "the pluralist's dilemma" was formulated, which means that a researcher needs to embrace different analytical perspectives while being able to express the reasons why her own approach is preferable to alternatives [84]. The discussion often focuses on Bayesian versus frequentist perspectives in statistics. It was stressed that the Bayesian approach might be used as an alternative to the frequentist one, so that researchers have "the right tool for the right job" [47]; instead of being claimed the only possible solution, each of these approaches can enrich the toolbox of suitable methods selected for a specific task [86].

Different ways of introducing method versatility can be distinguished based on literature, as shown in Figure 2.1. I labelled them the consecutive use, the toolbox choice, the simultaneous use, and the range extension. The list is not exhaustive, and a different categorisation can be as valid as the one presented here.

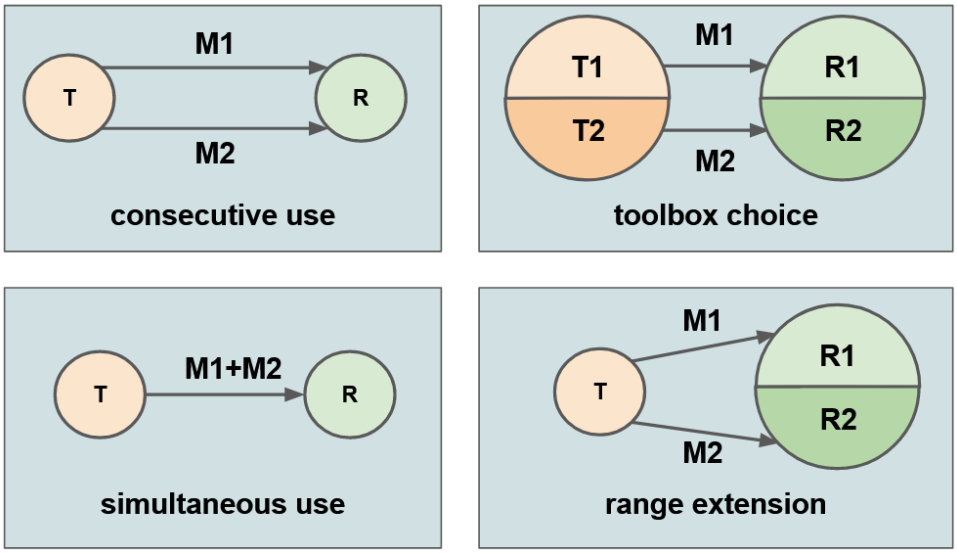

Figure 2.1: **Introducing Method Versatility.**
T stands for task, M for method, R for result.

Firstly, after using a default method, the researcher can reconfirm the results with a new method applied consecutively to the same dataset. I refer to this way of introducing method versatility as "the consecutive use" of different methods. In literature, it was also called the iterative process [220] implying that the same task can be consecutively readdressed from different perspectives. Thus, another (for instance, a novel, or a more advanced) method helps to reassess information obtained by a commonly used method [58].

Secondly, methods from different approaches can be flexibly chosen for different tasks. For instance, the toolbox approach [86] was suggested for Bayesian and frequentist statistics, so that neither is applied indiscriminately for any task. Approaches adopted by rapidly developing areas of educational data mining and learning analytics [211] also imply choosing a suitable instrument from statistical and ML toolboxes. Conceptually, when Gelman [84] maintains that the way out of the pluralist's dilemma is to recognize that different methods are appropriate for different problems, it refers to the toolbox approach.

Thirdly, different methods can be applied simultaneously and integrally to obtain new insights that would have not been possible with each of them used separately. For instance, such recently developed approaches as statistical learning [105], statistically reinforced ML [213], and bi-dimensional approach [229] suggested simultaneous use of statistical and ML methods, e.g., augmenting machine learning models by investigating significance and effect sizes typical for statistical models.

Fourthly, introduction of a new method can be useful in that it extends the range of possible approaches to a task. This can be a motivation for developing a new clustering algorithm, as the task of finding meaningful groups in the data can be handled in

various ways, considering the inevitable subjectivity of choices made in this process [64]. Novel methods of scholarly communication most frequently belong to this group, as the same task – to present the research findings to the academic community – is approached from a different perspective. For instance, it is the case for SKGs that improve machine readability and accessibility of the research results [13].

These ways of introducing method versatility can be used independently, or combined in different forms. For instance, a statistician might apply a "default method" (e.g., logistic regression for a classification task) and then reassess its results – thus exercising "the consecutive use" - with a technique of statistically reinforced ML, which itself can be viewed as "the simultaneous use" of different methods. This categorisation is therefore somewhat artificial, and only useful for the purpose of imposing a certain structure on a wide variability of methodological choices.

2.2 Methods of Psychometric Analysis

In this section, I discuss methods applied for assessing the main characteristics of a psychometric instrument, such as validity, dimensionality, and reliability. Item-level functioning, or item functioning, is explored to obtain a more detailed picture. Most of these methods belong to either CTT or IRT, which are well-established theoretical frameworks.

CTT is a long existing framework that is widely accepted in psychometric research. The measurement precision in CTT is assumed to be equal for all individuals. According to CTT, a "true score" of a participant is the expected score over an infinite number of independent administrations of the scale. Thus, an observed score consists of a true score and a component caused by a random measurement error:

$$X = T + E$$

Here, X is an observed score, T is a true score, and E is a measurement error [36].

IRT is a collection of probabilistic models that describe the relationship between the observed score and the underlying latent trait. The models have different numbers of parameters (one, two, and three-parameter models) and can deal with dichotomous or polytomous items [14]. The measurement precision in IRT depends on the latent trait value. IRT is consistent with a cognitive theory of how people respond to questions [231]; it allows maintaining the width of the latent continuum and diagnosing whether the test is able to differentiate between the respondents' ability on the latent dimension, which decreases the occurrence of Type II error [58]. Rash models belong to a widely used family of one, two- and three-parameter models that take principles of measurement from physics and emphasize invariance in measurement. Rating Scale Model (RSM) is a one-parameter Rash model for polytomous data that can be described by the following equation [261]:

$$ln\frac{P_{n,i(x=k)}}{P_{n,i(x=k-1)}} = \theta_n - \delta_i - \tau_k$$

Here, *P* is the probability that a person *n* will endorse the category *k* in a polytomous item *i*, θ is ability of the person *n* on the latent dimension, δ is the difficulty of the item *i*, and τ is the threshold for the category *k* of this item. Thresholds are the points on the logit scale at which there is an equal probability to endorse the category of interest or the category just below it. They are estimated empirically for all items. Probabilities of endorsement for items and categories can be visualized with item trace lines, which are also called item characteristic curves (see Figure 2.2).

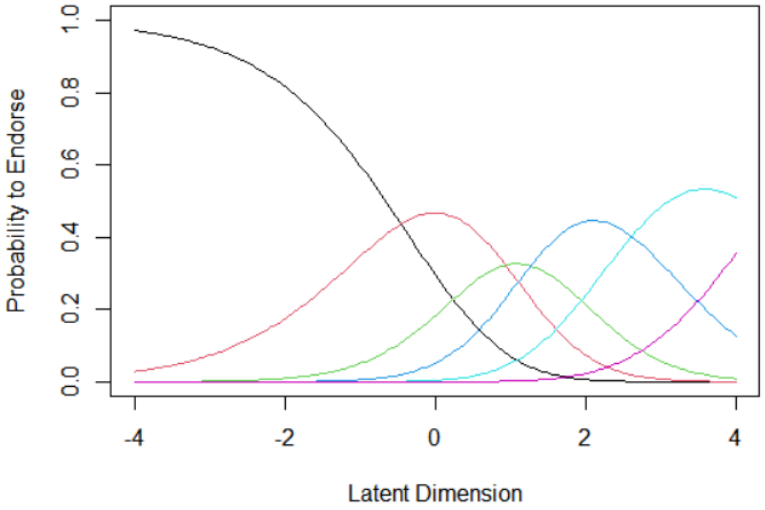

Figure 2.2: **Item Trace Lines for a Polytomous Item.**

Nonparametric IRT is a framework developed to relax some of the parametric IRT assumptions. It includes Mokken models, or Mokken Scaling Analysis (MSA). MSA evolved as a probabilistic version of Guttman scaling, which is deterministic and does not allow for randomness [258]. The advantage of MSA is that it relaxes strict assumptions on the shape of item characteristic curves typical for parametric IRT, and in particular, for Rash models [238]. Both parametric and non-parametric IRT are increasingly used in psychometric research.

A psychometric procedure in the frame of either of these frameworks, as well as any statistical test in general, can be suitable for some datasets but not for others. In particular, sample size considerations should be always taken into account [71], as a test conducted on an insufficient sample will lack statistical power, and inferences cannot be made. Another important and often neglected issue is assessment of distributional assumptions. For instance, many of CTT methods require the multivariate normal distribution of the data, which is a multidimensional generalisation of the unidimensional normal distribution. Failure to check the assumption might lead to an inadequate model that seemingly fits the data but gives biased estimates [93].

Therefore, checking assumptions plays an important role in analysing the psychometric properties of the scale, such as validity, reliability, dimensionality, and item functioning.

Validity is the crucial characteristic of a scale, without which other characteristics become rather inconsequential. However, methods of analysis and transparent reporting of this psychometric property have yet to be improved [113]. Construct validity, which consists of convergent and discriminant validity, is the extent to which an instrument assesses a construct of concern. Convergent validity is the extent to which the scores of the scale are related to scores of scales measuring theoretically similar constructs. Discriminant validity is the extent to which the scores of the scale are different from scores of scales measuring theoretically unrelated constructs. These relationships between the construct under scrutiny and other constructs can be presented as a nomological network. Typically, convergent validity is assessed by correlations between the scores on the scale under study and scores on existing measures for similar constructs, and discriminant validity by correlations between the scores on the scale and the scores on existing measures for conceptually different constructs [22]. Precise mathematical definitions of discriminant validity vary, as well as recommended methods of assessing it. To obtain more unbiased results on construct validity, more than one method is advisable [210].

Dimensionality is important, as the researcher needs to understand whether the scale is a unidimensional instrument, or if it consists of a few facets (subscales) measuring somewhat different aspects of the construct. The most common tool for studying dimensionality of psychometric instruments is factor analysis (FA), which can be exploratory (EFA) or confirmatory (CFA). FA reduces dimensions of the data to fewer latent variables while retaining as much information as possible, and thus it can be used to determine the number of factors that correspond to dimensions, or subscales, of the psychometric instrument.

Statistical and methodological decisions required by EFA were summarized in [116], and recommendations on rigorous practices were given. Prior to EFA, the data should be inspected to decide whether this method is appropriate. Assumptions for EFA can be checked with Barlett's test of sphericity and Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. Barlett's test checks the hypothesis that the observed correlation matrix is not an identity matrix, which would imply lack of relationships between variables in the dataset. KMO is an indicator of latent factors in the data as shown by common variance. It gives more detailed results than Barlett's test, which can be either significant or nonsignificant, and therefore, both procedures are recommended for detecting violations of the assumption. Then, a factor analytic method should be chosen. PCA is frequently applied for this purpose, but it is based on a different mathematical model than EFA techniques, and it does not account for the structure of correlations. Therefore, choosing Principal Axis Factoring (PAF) or maximum likelihood rather than PCA as a factor analytic method is recommended. For factor retention, the researchers are advised to use scree plot analysis in combination with

either parallel analysis, or Velicer's Minimum Average Partial (MAP) test, or both, rather than Kaiser criterion [116].

When the structure of a scale is supposed to be known from previous studies, it can be assessed with CFA. This method requires checking the multivariate normality assumption, which is often neglected in CFA studies [93]. CFA can be conducted with different estimators. Maximum Likelihood with Robust standard errors (MLR) estimator was shown to give sufficiently unbiased results; however, for a large sample, other estimators can be applied [168]. To assess the global fit between the tested model and the data, it is recommended to consider the results of the chi square test, which should reveal no significant differences between the model and the observed covariances; the Comparative Fit Index (CFI), which compares the fit of the model with the fit of a null model; the Root Mean Square Error of Approximation (RMSEA) as a parsimony-adjusted index; and the Standardized Root Mean Square Residual (SRMR), which is the square root of the difference between the residuals of the sample covariance matrix and the model [117]. These fit indices, as well as any modifications of the model, if there are valid reasons for them, should be reported thoroughly and transparently [124].

Dimensionality can also be explored with Very Simple Structure Analysis (VSS) and Item Cluster Analysis (ICLUST), though these methods are less common that FA [201]. VSS assesses the fit of several models of increasing complexity (that is, with the increasing number of factors) using the residual matrix of each model. ICLUST hierarchically clusters items of the scale and visualizes the results as a cluster diagram. The clustering approach is bottom-up, and items are grouped in subscales so that the value of internal consistency reliability of the resulting subscale is maximized [202]. The indicators of internal consistency reliability are described below, and hierarchical clustering of variables is discussed in more detail in section 3.2.

In the frame of IRT, dimensionality assessment can be conducted with MSA [238]. This approach does not imply the restrictive assumption of multivariate normality and thus gives less biased estimations of dimensionality compared to FA [58]. MSA includes checking the assumptions of homogeneity, monotonicity, local independence and invariant ordering. Homogeneity means that the items form a single scale measuring the same latent trait (θ), taking into account unique properties of the items and the measurement error. Monotonicity means that for each item, the probability of a particular response level Pi(θ) is a monotonically non-decreasing function of the latent trait θ. The local independence assumption states that one's response to an item *i* is not influenced by the responses to the other items of the scale. The invariant ordering assumption asserts that the items are ordered on their difficulty in the same (invariant) way at all levels of the latent dimension θ. The MSA procedures allow detecting items that violate the assumptions [238], and thus, at the stage of the scale development, a suitable set of items can be selected.

One of the most important characteristics of a psychometric instrument is reliability. Internal consistency reliability can be understood as the consistency of results given by

any item or group of items and the scale in general. Statistically speaking, it is the degree to which the items co-vary [71]. Internal consistency reliability is still most frequently estimated with Cronbach's alpha, which calculates the proportion of variance in the scale score explained by the trait being measured in relation to the total variance. This coefficient, however, is prone to bias, as the value of alpha increases with the number of items in the scale. In addition, Cronbach's alpha is sensitive to negatively formulated (reversely coded) items, and it is based on assumptions that are hardly ever met [60]. Other reliability coefficients include beta, a more conservative coefficient that explains variance in the data by a general factor via average covariance between items of the worst split half. The latter means such a split of the scale that minimizes the covariance value. Guttman's lambda-6 is an estimate based on the item variance accounted for by linear regression [58]. Currently, MacDonald's omega is considered the most unbiased estimate of internal consistency reliability. Similarly to Cronbach's alpha, it calculates the ratio of the trait-related variance to the total variance, but under more realistic assumptions. It was recommended to report the commonly used alpha and the more unbiased omega with confidence intervals, as these are more informative than point estimates [60].

Item-level functioning, or item analysis, explores individual items of the scale. Item difficulty and item discrimination are assessed differently in the frame of CTT and IRT. In CTT, item difficulty for a polytomous item is its mean value, and item discrimination is the corrected (that is, calculated with this item removed) item-to-scale correlation. In IRT, such as Rash models for polytomous items, item difficulty can be understood as the point on the latent continuum at which the highest and lowest categories have equal probability of being observed. In item trace line visualization, it refers to the centre of the middle category for an odd number of categories, or the transition between adjacent central categories for an even number of categories. Item properties in the frame of IRT are characterised by item fit and person fit measures [257]. Person fit indicates how many respondents have response patterns that do not fit the model. Person-item map shows the location of person abilities and item difficulties estimated by the model along the same latent dimension. This visualization can be used to explore the extent of item coverage, determine the comprehensiveness of the scale, and detect redundant items [36].

To summarise, psychometric properties of a scale can be assessed by methods, most of which belong to the frameworks of CTT or IRT. Validity, dimensionality, reliability, and item functioning of a scale can be explored in different ways, and researchers ought to widen the scope of their psychometric techniques to include versatile methods from both frameworks. Checking assumptions, which are required for psychometric tests to give unbiased results, could inform further analytical choices and is a prerequisite of methodologically rigorous research.

2.3 Methods of Educational Data Analysis

In this section, I outline the concepts related to educational data analysis, and specifically, data analysis of large-scale educational surveys. In educational science, studies with rather small samples are not infrequent, which are problematic in terms of statistical power of the tests and generalizability of their findings; large-scale educational surveys, however, are exempt from this shortcoming. In the frame of these surveys, teachers' and students' attitudes towards technology are measured among many other individual characteristics, learning achievements, skills, and demographic variables, such as Economic, Social, and Cultural Status (ESCS) and gender [63], [76], [77].

Two of these international surveys, PISA and ICILS, are selected for this work. PISA is conducted by the Organisation for Economic Cooperation and Development (OECD) once every three years to measure 15-year-old students' literacy (competence required for coping with adult life) in different domains, attitudes to schooling (including attitudes towards ICT), and a broad range of demographic factors. ICILS is conducted by the International Association for the Evaluation of Educational Achievement (IEA) to measure various aspects of students' and teachers' interaction with ICT. The data collected in the frame of these surveys are of high quality, and the samples are sufficiently large to be suitable for a wide spectrum of methods, which are discussed below. International comparisons in these surveys were criticised as insufficiently valid for a number of methodological reasons (see, for instance, [88] and [197]); therefore, single-country (German) samples are analysed in this work.

2.3.1 Supervised Learning

Supervised learning, which includes classification and regression tasks, relies upon techniques traditionally used in the frame of statistical analysis or developed more recently in the frame of ML [164]. ML differs from statistics in regard to data pre-processing (splitting the data into the training set and the test set), choice of variables (less theory-driven than in statistics), and model evaluation (predictive accuracy instead of explanatory power). ML strives to tackle the bias-variance trade-off in order to achieve high predictive accuracy, while statistics focuses on model estimation, inference and fit [229]. In section 3.1, attempts to integrate these approaches are discussed, and in this section, I describe methods used in the frame of each of them and relevant to the tasks of this thesis. In particular, I focus on the RF algorithm for supervised learning tasks and missing data imputation; describe statistical analysis with plausible values and replicated weights; and outline methodological advantages and pitfalls of HLM.

Machine Learning: Random Forest

RF belongs to the family of algorithms based on decision trees. The idea of a decision tree is to split the dataset on a particular variable based on information gained from this

split. As splits are binary (yes or no), continuous variables are transformed into categorical variables, that is, lesser than or greater than a certain value. For illustration purposes, I built a decision tree with the PISA data to predict students' mathematical proficiency (the classification task) based on their attitudes towards ICT, ESCS, and gender. The tree is shown in Figure 2.3. Node labels contain the following information: the predicted class for the node; the probability per class of observations in the node (conditioned on the node, the sum across a node equals one); and the percentage of observations in the node. For this decision tree, the first two splits are based on the ESCS value, and the third on the value of an attitude towards ICT (a student's ICT autonomy). Split labels specify how the decision tree splits the data.

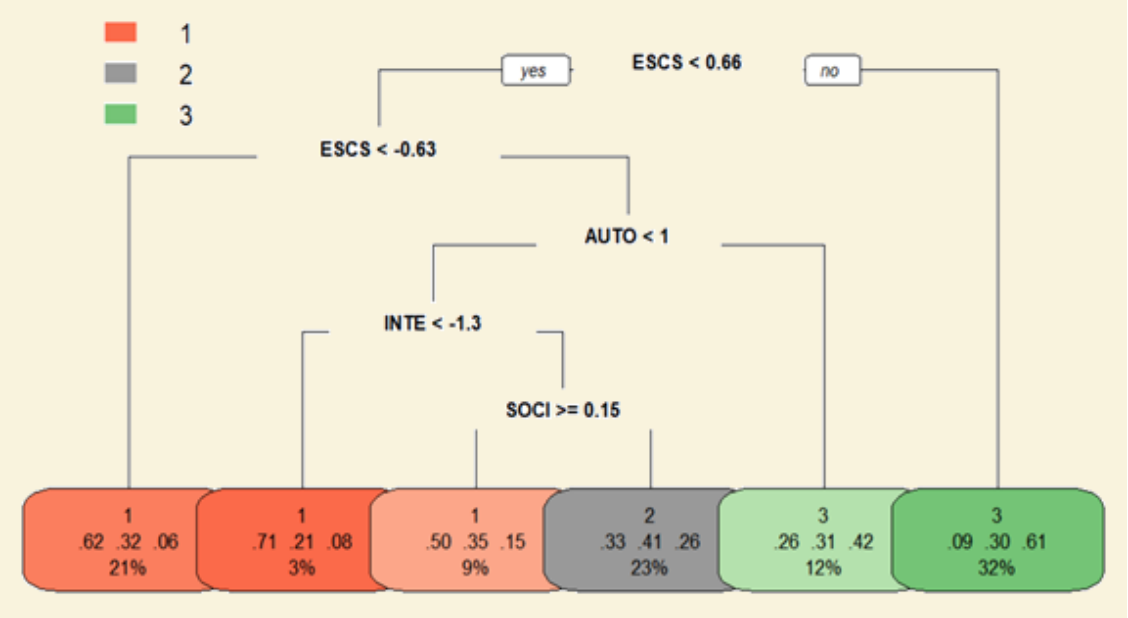

Figure 2.3: **Decision Tree for Students' Mathematical Proficiency Levels.**
ESCS is economic, social, and cultural status; AUTO is ICT autonomy; INTE is ICT interest; SOCI is ICT in social interaction; class 1 is mathematical proficiency below Level 2; class 2 is Levels 2–4; class 3 is Level 5 and above.

Decision trees have a number of advantages: they are interpretable as they mirror human decision-making more accurately than other predictive models, they handle a wide range of problems and different types of variables, and they impose no assumptions on the data [90]. Their substantial disadvantage is the problem of overfitting, when the algorithm is unable to generalize well to new data.

This problem is resolved in RF by generating a large number of bootstrapped trees (a 'forest' of decision trees) based on random samples of variables and aggregating their results. RF has gradually grown from a single algorithm [28] into a framework of various models [55], [69]. RF is effective for highly dimensional data and handles interactions and nonlinearity [69], [164]. It has become a widely used method for classification tasks [55], as it has lower predictive error than logistic regression [29],

[50]. For multiclass tasks, it does not require the proportional odds assumption needed by ordinal regression [71], [172]. RF is increasingly used for missing data imputation, as it is adaptive to interactions and nonlinearity [227], performs well even with data missing not at random [245], and has better imputation accuracy than nearest neighbours imputation and multiple imputation by chained equations [163], [256]. To measure model performance of RF, different metrics can be used [41], [125]. One of the most commonly used metrics is the area under the Receiver Operating Characteristic (ROC) curve [132]. ROC is a probability curve, which is plotted with true positive rate on the y-axis against false positive rate on the x-axis. The area under the curve (AUC) represents the ability of the model to separate classes. For a perfect classifier, the area is 100%, and for a random classifier, when true positive rate is equal to false positive rate, it is 50%. The AUC can be extended to multiclass problems, with separate comparisons for each pair of classes [101].

RF models are better predictors than decision trees, but they are less understandable. Breiman [29] described this trade-off between accuracy and interpretability of a model as the Occam dilemma. In order to make the output of such 'black boxes' more understandable, model-agnostic (that is, flexibly applicable to any model) methods can be used, such as permutation variable importance and partial dependence plots [164]. The concept of permutation variable importance can be briefly explained as follows. A variable is considered important as much as deleting it would affect prediction accuracy of the model [29]. However, removing one variable after another and retraining the model each time would be computationally expensive, and instead of it, we replace a variable with random noise to see how it influences the model's performance. The noise is created by permuting the variable (that is, shuffling its values). Permutation variable importance was shown to be less biased than other variable importance measures [5], [240].

Partial dependence plot for a variable shows the marginal effect that it has on the predicted outcome of the model [164]. It depicts the nature of the relationship between the variable and the outcome and indicates whether this relationship is linear, monotonic etc. In classification tasks, partial dependence plots can be built for each class separately to show the probability for the class given the values of the variable. Partial dependence plots can be constructed for two variables in 3D to display not only their influence on the outcome but also their interactions with each other [94].

Based on the findings showing effectiveness of RF for missing data imputation and for classification, it can be concluded that the RF algorithm is a suitable method for these tasks. It can be combined with model-agnostic methods (variable importance and partial dependence plots) to increase interpretability of the models.

Statistical Methods for Educational Surveys

In large-scale educational surveys, data are often hierarchical. It means that some variables are nested within other variables: for instance, students in one school can have similarities with each other and differ from students in the other school, so that

school-related differences are as influential as individual differences, or even more. Thus, a grouping variable (school) should be included in the model, and other school-level context variables might be relevant as well.

In Figure 2.4, an example of hierarchical data from PISA study is presented. For students of different schools, the relationships between their interest to ICT and ESCS vary, so that the regression lines have different slopes and different intercepts.

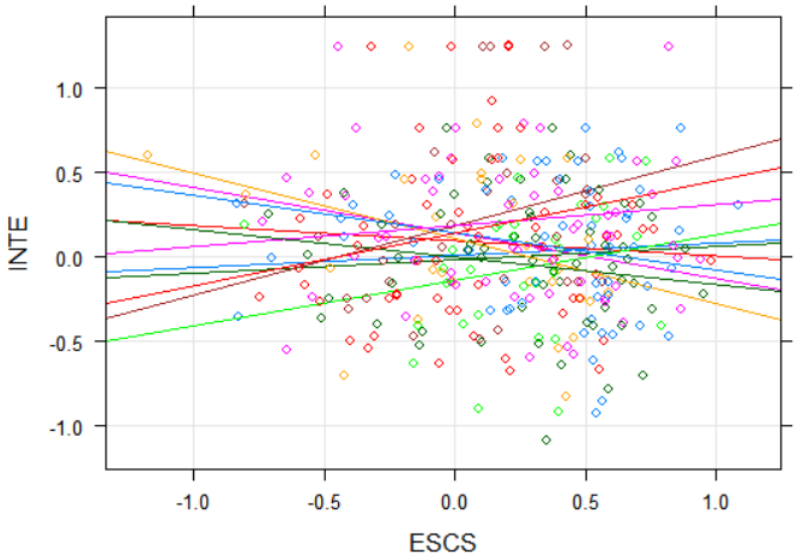

Figure 2.4: **Hierarchical Data: Varying Slopes and Intercepts**.
ESCS is economic, social, and cultural status; INTE is ICT interest.

For hierarchical data, multilevel models are most appropriate [178] because they take into account the structure of the data, and failing to recognize hierarchical structure of the data leads to Type I error inflation [165]. It was shown that multilevel approaches work better than fixing standard errors of ordinary least squares estimate [44]. Multilevel approaches include multilevel structural equation modelling (SEM) [11], [89], multilevel latent class analysis (LCA) [265], multilevel multidimensional IRT [149], and the most frequently used method, HLM.

HLM is a linear regression model for hierarchical data that allows for variability in intercepts and slopes. Therefore, this model does not require the assumption of homogeneity of regression slopes, or independence of different observations; it explicitly models non-homogenous slopes and relationships between cases in the data. However, distributional assumptions for residuals should be checked thoroughly [71], and special attention should be paid to the assumption of non-collinearity of predictors. This assumption can be checked by calculating values of the variance inflation factor, which indicates the strength of correlations between predictors; the cut-off value of 3 was recommended as acceptable [273], meaning that multicollinearity is not present in the data.

Distributional assumptions for residuals can be easily checked with diagnostic plots. These include the normal distribution of residuals and the homoscedasticity assumption, which means that the variance of residuals should be equal across different levels of predictors. Figure 2.5 shows examples of diagnostic plots for HLM, which I built with the package sjPlot [150] to illustrate assumptions about residuals.

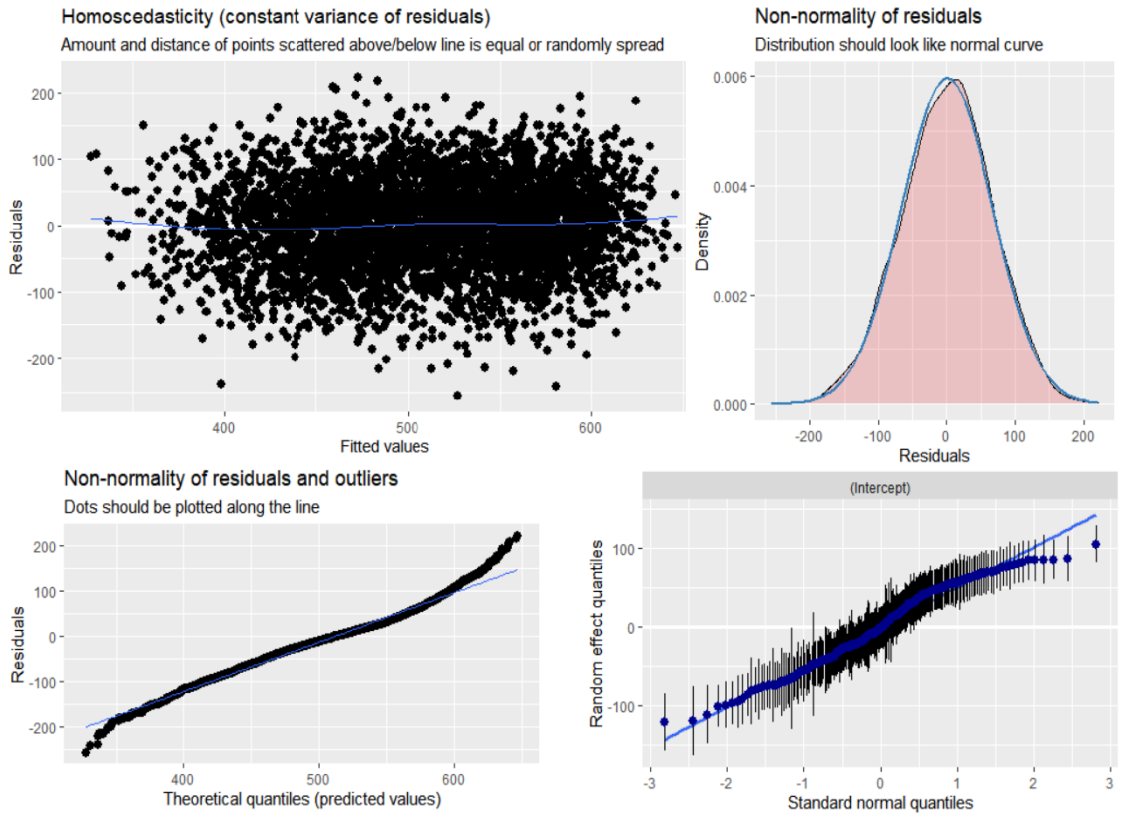

Figure 2.5: **Assumptions for HLM with Diagnostic Plots.**

With the development of multilevel modeling, possible pitfalls in application of HLM were recognized, and guidelines for overcoming them were elaborated [3], [160]. According to the methodological guidelines, models with random intercepts and random slopes should be fit whenever possible; in case these models fail to converge, models with random intercepts can be explored. In terms of data preprocessing, grand-mean (rather than group-mean) centering of independent variables was recommended [103]. Variable selection (feature selection) methods for multilevel models were outlined in [107]. The standardization by two standard deviations was suggested, as it makes comparisons between continuous and binary variables possible [83]. An unconditional model should be compared with other models, with intraclass correlation coefficients (ICCs) and proportional reduction of variance reported [103]. Effect size measures for multilevel models were discussed in detail in [138] and [147]. Generally, Nakagawa's marginal and conditional R^2 (see [166], [167]) can be reported as effect size measures for random effects. When the researcher follows these recommendations,

HLM can be an effective way of dealing with the hierarchically structured data typical for large-scale educational surveys.

As it is often the case with large-scale surveys, analysis of PISA data should include plausible values and replicate weights. Plausible values are random values from the posterior distributions. Replicate weights are more informed standard error estimates retaining information about the complex sample design. Instructions given by PISA Data Analysis Manual [178] on computation of statistical estimates and standard errors with plausible values and replicate weights can be summarized as follows:

To calculate a final estimate (e.g., a regression coefficient), estimates for models with all plausible values and the final weight are averaged ([178] , p. 118):

$$\theta = \frac{1}{M}\sum_{i=1}^{M}\theta_{\mathrm{i}}$$

The standard error of this final estimate is calculated as follows:

The sampling variance is calculated by averaging sampling variances for all plausible values ([178], p. 118):

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}{\sigma_{\mathrm{i}}}^2$$

Each of them is calculated as the averaged (with the coefficient specified below) sum of squared differences between the final estimate and each of replicate estimates ([178], pp. 73–74):

$${\sigma_{\mathrm{i}}}^2 = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}(\theta_{\mathrm{i}} - \theta)^2$$

The imputation variance is calculated as the averaged (with the coefficient specified below) sum of squared differences between the final estimate and an estimate for each of plausible values ([178], p. 100):

$$B = \frac{1}{M-1}\sum_{i=1}^{M}(\theta_{\mathrm{i}} - \theta)^2$$

The final variance is calculated as the sum (with the coefficient specified below) of the sampling variance and the imputation variance ([178], p. 100):

$$V = \sigma^2 + \left(1 + \frac{1}{M}\right)B$$

The standard error of the final estimate is the square root of the final variance ([178], p. 119). In these formulae, θ is a statistical estimate (e.g., a regression coefficient); σ^2 is the sampling variance; *B* is the imputation variance; *M* is the number of plausible values (for the 2015 and 2018 datasets, 10); *G* is the number of replicate weights (for the 2015 and 2018 datasets, 80); *k* is a deflation factor. PISA uses the Fay method with a *k* factor of 0.5 ([178], p. 73). This approach allows making unbiased estimates of regression coefficients and their standard errors in regression models, such as HLM, when these are applied to PISA data.

To summarise, classification and regression tasks in educational data analysis can be handled by ML and statistical methods. For large-scale educational surveys, such as PISA, statistical analysis requires dealing with plausible values and replicate weights. HLM is a statistical tool of choice applied to hierarchical data, and the RF algorithm is one of the most effective and unbiased methods for supervised learning tasks and missing data imputation. Based on these findings reported in literature, the flexible use of ML and statistical methods for data analysis of large-scale educational surveys can be feasible.

2.3.2 Unsupervised Learning

Unsupervised learning deals with unlabelled data. An example of such approach is clustering, which is discussed in detail in this section. This work deals with model-based clustering, specifically, Latent Class Cluster Analysis (LCCA), and the problem of selecting the number of clusters. As clustering discerns latent groups of observations in the data, it is categorized as a person-centred approach [263] and is extensively used in organizational psychology and educational science. In particular, LCCA was applied to the teachers' data from ICILS 2013 [63]. Here, basic concepts related to distance-based and model-based clustering and different approaches to selecting the number of clusters are explained.

Distance-Based and Model-Based Clustering

Clustering evolved as an unsupervised ML technique [23] aimed at finding similarities between observations. It is a rapidly changing area, in which novel algorithms constantly arrive [199]. Performance of clustering algorithms can be compared by means of internal criteria, which assess the result itself [99], or external criteria, which compare it to a reference result, e.g., the known cluster partition [253]; in some studies, both internal and external criteria are used [209].

Various taxonomies of clustering methods can be found in literature (see, for instance, [4], [64], [263]); in the frame of this work, distance-based and model-based clustering [7], [199] are discussed. Distance-based methods, which could be also called dissimilarity-based [111] or partitional [4], belong to a wider group of methods that were labelled as heuristic [95] or algorithmic [263]. Distance-based methods conduct partition of observations based on a dissimilarity criterion, while model-based methods

fit probabilistic models to the data; the former are mostly ML techniques, while the latter were developed in the frame of statistical analysis [64].

Distance-based clustering includes unsupervised learning methods that are simple and computationally inexpensive [23]. These methods require that the number of clusters is chosen prior to the analysis. The most widely used distance-based method for interval data is k-means [189], which was first introduced in 1960s [155]. The algorithm starts with random assignment of each cluster centre, or centroid, and observations are clustered to the nearest centroid; then centroids are repositioned based on the created clustering. The process of relocation of centroids and re-clustering of observations repeats until a stable configuration is found (the cluster centroids do not move). For categorical data, k-medoids and k-modes are extensions of k-means. In k-medoids, or partition around medoids, an actual central data point is used as a measure of the centre instead of the cluster centroid, while in k-modes, clusters are defined based on the number of matching categories between data points [42].

Model-based clustering, which is also described as finite mixture modeling [159], started as latent structure analysis [141] implemented with the expectation maximization algorithm [54]. Model-based methods for the interval data, assumingly represented by a mixture of normal distributions, are called Gaussian mixture models, or profile analysis [228]. For categorical variables with assumed multinomial distributions, the model-based method is LCCA [141]. It can be applied to dichotomous [30] and polytomous categorical data. In the frame of statistical analysis, LCCA (also termed LCA) is defined as a statistical approach to modeling a discrete latent variable using multiple, discrete observed variables as indicators [140]. This approach allows to the local independence assumption to be relaxed [251], covariates to be included in the model (see [235], [250]), and the bootstrap likelihood ratio test to be conducted for the model selection [61], [170]. In this work, LCCA is considered exclusively from the clustering perspective, so that the discrete latent variable represents the cluster assignment. Previous studies showed that LCCA was an effective method for clustering categorical data, while a “naive” approach (treating the data as interval and applying k-means) resulted in poor performance [199]. As a more flexible model-based clustering tool, LCCA is preferable in many real-world circumstances, e.g., unequal covariance matrices, unequal numbers of observations in clusters and poorly separated clusters [7]. LCCA represents the next generation of tools that provides the researcher with the wealth of diagnostic information, and therefore it is increasingly used in educational research [32], [63], [65], [146].

Selecting the Number of Clusters

Selecting the number of clusters was discussed in literature as a rather controversial [95] and philosophical [262] topic. It was emphasized that clustering is always “in the eye of the beholder” [64], and it would be meaningless to speak about unique objective “true” clusters [111]. In practical terms, the number of clusters in any clustering method is selected based on pre-specified criteria. The most commonly used metrics for

distance-based methods are the elbow method for the within sum of squares, the gap statistics, and the Average Silhouette Width (ASW) [23]. For model-based approaches, in which clusters are understood as latent classes, information criteria are used. When there is no agreement between different criteria, it might be difficult to select the number of clusters [205], and domain knowledge of the researcher plays a pivotal role [269]. The interpretability of clusters, parsimony of the solution, and the size of population shares should be taken into consideration [190].

Here, I describe approaches and criteria used in this work for selecting the number of clusters in LCCA. They are related to model fit, cluster separation, and the stability of cluster partitions.

Model Fit. The Bayesian Information Criterion (BIC), which was introduced by Schwartz [226] in 1970s, is one of the most well- performing and widely used information criteria. It is defined as follows:

$$BIC = -2 \log L + p \log n$$

where p is the number of free parameters in the model, n is the number of observations, and L is the maximized likelihood function of the model. For a large n, minimizing the BIC corresponds to maximizing the posterior model probability [262].

In addition to the BIC, the Integrated Completed Likelihood (ICL) criterion was introduced by Biernacki et al. [21]:

$$ICL(m, K) = \max_{\theta} \log f(x, \tilde{z} | m, K, \theta) - \frac{\upsilon_{m,K}}{2} \log n$$

where x is the data, $\tilde{z}$ is the estimated cluster membership for observations in the model m with K as the number of clusters, θ refers to the estimated mixture parameters, and $\upsilon_{m,K}$ is the number of free parameters in the model. The ICL is equal to the BIC penalized by the estimated mean entropy [19], which means that it aims at finding well-separated clusters and thus should not overestimate the number of clusters [95].

Cluster Separation. The ASW is the criterion traditionally used for selecting the number of clusters in distance-based methods. The ASW is the averaged value of silhouette widths for observations, which are defined as follows [212]:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

where $a(i)$ is average dissimilarity between observation i and all other points of the cluster to which i belongs, and $b(i)$ is average dissimilarity between i and all observations of the nearest cluster to which i does not belong. To calculate the ASW, a dissimilarity (or distance) measure should be selected [254]. For categorical variables, the dissimilarity measure defined by Kaufman and Rousseeuw [130] is typically used, as it was shown that none of the dissimilarity measures for categorical variables is

always superior or inferior to others [24]. The dissimilarity between two rows is thus calculated as follows:

$$d_{i,j} = \frac{\sum_{k=1}^{p} \omega_k \delta_{ij}^{(k)} d_{ij}^{(k)}}{\sum_{k=1}^{p} \omega_k \delta_{ij}^{(k)}}$$

where the dissimilarity $d_{i,j}$ is the weighted mean of the contributions of each variable $d_{ij}^{(k)}$ with weights $\omega_k\delta_{ij}^{(k)}$. When weights $\omega_k\delta_{ij}^{(k)}$ are not specified, they are equal to 1. The contribution $d_{ij}^{(k)}$ of a variable to the total dissimilarity is 0 if both values are equal, and 1 otherwise. The ASW values range from –1 to 1, and higher positive values indicate better defined clusters characterized by within-cluster compactness and between-cluster separation, while values close to 0 or negative values show that the clusters are not well-separated [48].

Figure 2.6 shows clusters with different separation. In the figure, multidimensional data is presented in two-dimensional projections, which are used here solely for illustration purpose; the ASW values and silhouette plots give a more precise picture of cluster separation.

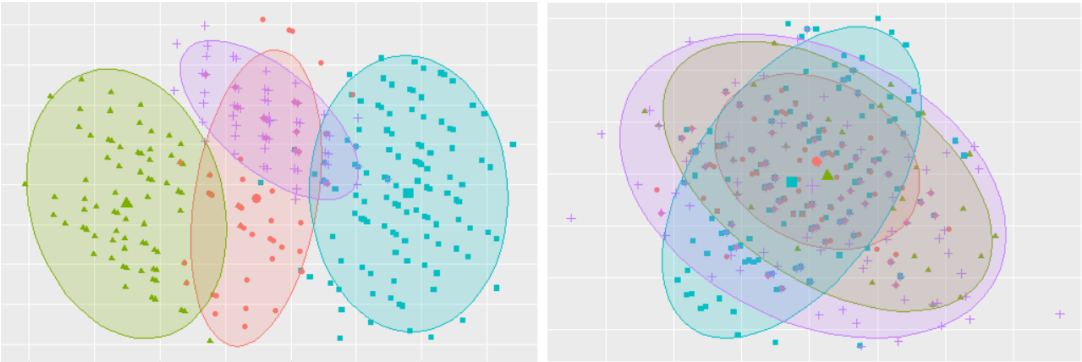

Figure 2.6: **Clusters with Different Separation.**

Stability of Partitions. Cluster validation should be conducted in order to evaluate the stability of clustering. For this purpose, bootstrap validation is typically used to check whether the chosen cluster solution depends on the specific dataset or can be generalised to the new data [23]. To perform this validation, we cluster the original data and apply the cluster solution to a bootstrap sample, which is also clustered anew. Thus, we have two cluster partitions for each bootstrap sample: the partition created by the original solution on the new sample and the new partition of this sample. They are compared using an external metric of our choice; this value is averaged over multiple repetitions to obtain the indicator of stability [110].

To compare partitions, external measures should be used, such as the Adjusted Rand Index (ARI) and the Jaccard coefficient [95]. These measures can be explained as follows [209]. We need to compare two different cluster partitions $U = \{U_1, U_2, \ldots U_r\}$ and $V = \{V_1, V_2, \ldots V_s\}$ conducted on the same data. Let n be the total number of

observations, and n_{ij} the number of objects in common between two partitions U_i and V_j, which sums as $n_{i.} = \sum_j n_{ij}$ and $n_{.j} = \sum_i n_{ij}$. There will be pairs of observations placed in the same cluster in both partitions:

$$a = \sum_{i,j} \binom{n_{ij}}{2}$$

Other pairs of observations will be placed in the same cluster in one partition but in different clusters in the other:

$$b = \sum_{i} \binom{n_{i.}}{2} - \sum_{i,j} \binom{n_{ij}}{2}$$

Still other pairs of observations will be in different clusters in both partitions:

$$c = \sum_{j} \binom{n_{.j}}{2} - \sum_{i,j} \binom{n_{ij}}{2}$$

.

The Jaccard coefficient is defined as

$$J = \frac{a}{a + b + c}$$

The ARI is defined as

$$ARI = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}$$

The ARI is one of the most frequently used external criteria [253]. The Jaccard coefficient is also widely used and easily interpretable as the proportion of observations placed in the same cluster in both partitions [109]. Both metrics assess the stability of partitions in bootstrap validation, with the higher value indicating the higher stability. In this work, the ARI and the Jaccard coefficient are used to find the number of clusters which maximizes the stability of partitions.

To summarise, selecting the number of clusters is a challenging task that requires a flexible approach. For that purpose, distance-based and model-based clustering apply different criteria. Model fit indices, such as the BIC and the ICL, and distance-based

measures such as the ASW, are widely used in cluster selection. The stability of cluster partitions should be taken into account for discerning generalizable clusters in the data. This work relies on these measures in selecting the number of clusters in LCCA.

2.4 Methods of Scholarly Communication: Knowledge Graphs

Scholarly communication in the XXI century is gradually shifting from the outdated document-centred to more advanced modalities. Currently, research findings are still shared with academic community via “pseudo-digitized” publishing formats that can be accessed electronically but are static and unstructured. These formats hinder machine readability of publications and eventually lead to a decrease in the reproducibility of research results [13]. KGs, as effective tools of information retrieval [200], are changing the landscape of information flows. According to a definition in [26], a KG consists of an ontology describing a conceptual model (e.g. with classes, relation types, and axioms), and the corresponding instance data following the constraints posed by the ontology. Nowadays, KGs are applied in various domains [272], including physics [218], healthcare [236], [270], business [161], and education [194], [206]. In academia, SKGs are novel scholarly communication instruments in the frame of the scholarly knowledge ecosystem [6]. SKGs deal with bibliographic metadata (e.g., [246]) and present research findings to the academic community in a more effectively structured machine readable form. The ORKG is an SKG that implements a research contribution model [252] encompassing actual results (contributions) from the academic literature. The ORKG research service infrastructure initiative integrates crowdsourcing and automated techniques for generating scholarly knowledge graphs [126] which enable the user to compare research contributions [174] and create FAIR literature surveys [175]. This cutting-edge technology is crucial to resolving the problems of contemporary scholarly communication.

The topic of wider acceptance of SKGs by the academic community was discussed in the frame of the ORKG [13] and, more generally, in KG research. A stronger culture of user-centric research with an interdisciplinary approach is required to involve specialists of different domains in the use of these novel instruments [215]. In particular, in order to facilitate the acceptance of SKGs by different professional and demographic communities, it is crucial to take into account mechanisms which humans employ to process information [192]. This would help to create SKG-powered interfaces appealing to various groups of users.

Effective methods of optimising information load, which take into account mechanisms of human perception, were suggested by cognitive load theory [243]. Cognitive load is a measure of the cognitive strain an individual feels when dealing with a task; for instance, in the context of user satisfaction with a website, cognitive load was defined as the amount of cognitive processing a person applies to find information [118]. Cognitive load can be intrinsic (that is, inherently present in a perceived object), germane (intertwined with the former and required to learn), and

extraneous (caused by the manner in which the information is presented). Extraneous cognitive load can be minimized by effective presentation of information. It was shown that that cognitive load has greater influence on user satisfaction with a website than performance outcomes; thus, reducing cognitive costs of information processing might be more important than increasing benefits [118]. Cognitive load can be preliminary assessed prior to being experienced by users, so that the design features can be planned holistically while taking into account essential user characteristics. To decrease extraneous load by effective structuring of information, a few principles should be taken into account that were confirmed by a few decades of research related to the theory [243]. The modality principle requires supplementing verbal material with visual and/or audial presentation of information. According to the spatial contiguity principle, visual and verbal information should be spatially integrated to avoid the split attention effect, when the user has to put cognitive effort into integrating disparate pieces of information. Presented information should not be redundant, to avoid the redundancy effect [37], and the most important information should be emphasised, to evoke the signalling effect, which can be also described as high affordance [66].

The crucial role of visual materials outlined by cognitive load theory is coherent with research findings. The visual modality “worth a thousand words” [264] and is beneficial in various contexts, such as instructional design [37]. Visual presentation of information schematizes relationships between the data, assists in establishing links between entities, elucidates similarities and differences between phenomena, facilitates pattern recognition, and supports understanding by saving cognitive resource of the user [46]. For information visualisation, researchers resort to such easy to use instruments as dashboards [188], which were applied in different areas for summarising and visually presenting data [9], [40], [67]. A dashboard can be defined as a visual display of the most important information needed to achieve one or more objectives, consolidated on a single screen [70]. It is an interactive tool with dynamically updated data that allows information monitoring [106]. SKG-based dashboards are yet to be developed in different domains, and this work takes an opportunity to extend the range of scholarly communication means and facilitate SKG acceptance in academia.

To summarise, SKGs as novel instruments of information retrieval play an important role in the transition from previous outdated to contemporary formats of communication. For wider acceptance of this technology, user-friendly SKG-powered interfaces need to be developed, which are based on principles of human information processing. The visual modality of presenting information implemented in dashboards might be beneficial in this regard, as it increases the ease-of-use of the technology and thus facilitates its acceptance.

2.5 The Topic of Human Attitudes towards Technology

This section discusses the topic of human attitudes towards technology, one of the most important topics to be measured, researched, and communicated in the contemporary world. An attitude is a multifaceted structure comprised of affective, cognitive, and

behavioural components [97]. Human attitudes towards technology can be conceptualized in different ways [268]; they are vitally important in our digitalised world, as they influence various areas of study, work, organisational behaviour, and individual well-being. This topic is the common thread of the thesis contributions. Each of these contributions focuses on a specific attitude or a set of attitudes described in this section.

Psychometric instruments are needed to measure human attitudes towards technology, and in the area of human-technology interaction, many such instruments were developed. However, some of them ignore individual differences [221], while others describe users' interaction with already outdated technology without taking into account the rapidly changing digital environment [12]. Franke, Attig, and Wessel [80] developed the ATI scale, an economic nine-item instrument that measures affinity for technology interaction. The authors defined affinity for technology interaction as the tendency to actively engage in intensive technology interaction. This construct is rooted in need for cognition, "a stable individual difference in people's tendency to engage in and enjoy effortful cognitive activity" [34], which is important for human information processing [39]. Affinity for technology had been explored by other researchers [62], [129], but the scales they developed were not sufficiently effective in terms of construct definition or dimensionality. In this thesis, the ATI scale was selected for psychometric reassessment, as it is a unidimensional economic scale measuring a clearly defined construct rooted in an established psychological attribute. The ATI items can be found at https://ati-scale.org/.

While the ATI scale is intended for a wide audience and captures interaction with a wide range of technologies, large-scale educational surveys are usually more specific in regard to the target group and the technology. For educational data analysis, I selected surveys that provide researchers with a large amount of high quality data, PISA and ICILS. In both cases, attitudes towards ICT in learning context are the topic of research; in PISA, these are students' attitudes, and in ICILS, I selected teachers' attitudes. They are based on different conceptual frameworks which I describe in this section.

In PISA, the ICT questionnaire changed with years. PISA 2003, 2006, and 2009 assessed students' confidence related to three types of ICT tasks (basic, Internet-related, and high-level tasks). In PISA 2012, positive components of attitudes towards ICT (perceiving ICT as a useful tool) were measured as a construct independent of negative components (perceiving ICT as an uncontrollable entity). In PISA 2015, a new conceptualization of ICT engagement [87] was introduced, and was also used in PISA 2018. This framework was based on self-determination theory [52] and included dimensions of ICT competence, ICT interest, ICT autonomy, and ICT in social interaction. The items of student ICT familiarity questionnaire, with the attitudes towards ICT covered by items IC013-IC016, can be found at https://www.oecd.org/pisa/data/2015database/. Scale indices for each dimension were obtained in the frame of IRT by means of generalized partial credit model, which

allowed for the item discrimination to vary between items within any given scale [182]. The scales are valid reliable instruments according to psychometric assessment [135], [162].

In ICILS, the teacher survey was designed as auxiliary to the student survey, and therefore, the topic was less highlighted in literature (see, for instance, [76] and [77]). ICILS 2018 studied, among other factors, teachers' positive and negative views on ICT. The items can be found at https://www.iea.nl/data-tools/repository/icils (to retrieve the data from the IEA website, it is necessary to agree to the terms and conditions associated with their use). To measure positive views on ICT, teachers were asked, to what extent they agree or disagree that using ICT at school helps students develop greater interest in learning, to work at a level appropriate to their learning needs, to develop problem solving skills, and facilitates other important aspects of learning. To measure negative views on ICT, teachers were asked, to what extent they agree or disagree that using ICT at school distracts students from learning, results in poorer calculation skills, and leads to other negative consequences. It was shown that attitudes to ICT were important factors of teachers' interaction with technology in the professional context [63].

Results of research on human attitudes towards technology can be shared with the academic community via means of scholarly communication, as discussed in the previous section. However, human attitudes towards technology are not only a possible topic of scholarly communication but also an intrinsic part of it, as they underlie the perception of communication means, which nowadays are predominantly digitalised. The degree to which researchers are eager to use a novel technology, including SKGs, or their behavioural intention to use the technology, depends on a number of attitudinal factors, which was studied in various domains. According to the usability paradigm, as exemplified by technology acceptance model, interaction with technology is determined by its perceived usefulness and perceived ease-of-use. Three decades of research have brought new insights to this paradigm; in particular, it was shown that perceived ease-of-use is a more influential factor for actual user experience than perceived usefulness [137]. In the frame of this paradigm, eight most common models of user acceptance, including the technology acceptance model, were integrated to develop the unified theory of acceptance and use of technology. According to the theory, there are four factors influencing the actual use of a technology. External factors are social influence and facilitating conditions, and internal factors are performance expectancy, which is an equivalent of perceived usefulness, and effort expectancy, which is an equivalent of perceived ease-of-use. Empirical research applying the theory to different contexts confirmed the key role of these factors in technology acceptance [249]. In terms of information theory, ease-of-use is related to optimally structured information aimed at avoiding information overload, as the latter is associated with information anxiety and information avoidance [232]. Excessively complicated, as well as oversimplified, input leads to decreased interest and reduced attention, thus stirring boredom [244].

The user experience paradigm emerged as a countermovement to the usability paradigm, as it focused on affective aspects of human-technology interaction beyond the task-related view [144]. In addition to pragmatic or instrumental qualities of technology, user experience models embrace its holistic, aesthetic and hedonic aspects [104]. Pragmatic qualities include perceived practicality and ease-of-use, while hedonic qualities are related to pleasure-bringing stimulation. Pragmatic qualities were found to be more influential for the outcome than hedonic qualities; in particular, perceived ease-of-use was shown to be a predictor of perceived usefulness, enjoyment, and behavioural intention to use the technology [115]. Therefore, perceived ease-of-use of technology positively influences users' attitudes to this technology and thus facilitates its acceptance.

To summarise, human attitudes towards technology is an important topic, as in our digitalised world they influence virtually every sphere of life. In this work, a few frameworks related to this topic are dealt with. These are the ATI scale as a psychometric instrument measuring affinity for technology interaction, the ICT engagement framework in PISA 2015 and 2018, and the ICT framework used in the teacher survey in ICILS 2018. Human attitudes towards technology could be a topic of scholarly communication, but more importantly, they need to be taken into account in developing contemporary means of scholarly communication, such as SKG-powered interfaces.

2.6 Summary

In this chapter, I outlined basic theoretical concepts and explained methods that are used in this thesis. In regard to method versatility, I schematized four possible ways to facilitate its introduction: the consecutive use, the toolbox choice, the simultaneous use, and the range extension, and gave examples of these approaches to method versatility implemented in previous studies. Then, I described the basic notions from the area of psychometric analysis. They include CTT and IRT frameworks, with parametric IRT (RSM) and nonparametric IRT (MSA) as distinct approaches. I outlined concepts of convergent and discriminant validity, dimensionality, internal consistency reliability, and item functioning, and explained how these can be assessed by methods of CTT and IRT.

The next section gave an overview of ML and statistical techniques applied to large-scale educational surveys. There was a specific focus on the RF algorithm as a ML method, HLM as a statistical method, as well as PISA data analysis with plausible values and replicated weights. In regard to unsupervised learning methods used for educational data analysis, I described distance-based and model-based clustering and gave an overview of methods that can be used for selecting the number of clusters in various clustering approaches: model fit, cluster separation, and the stability of partitions.

In relation to scholarly communication, I briefly described SKGs as the novel modality of information retrieval and explained the main principles of technology acceptance that should be taken into consideration for developing user-friendly interfaces to facilitate the use of SKGs by a wider audience. The topic of technology acceptance is reiterated in the last section of this chapter, in which I discussed human attitudes towards technology and the ways these are measured, researched, and communicated.

Chapter 3

Related Work

In this chapter, I summarise the current state of research in the areas related to the thesis. I start with describing novel methods of psychometric instruments validation, which include an analytic procedure combining CTT and IRT techniques, and methods of assessing construct validity. In regard to supervised learning, I describe ML and statistical methods recently applied to large-scale educational surveys and give an overview of studies suggesting integration of statistical and ML approaches. In regard to unsupervised learning, I refer to previous works that criticize cluster selection based solely on a model fit index and explore the possibility of an integral approach to selecting the number of clusters. In the last section, I outline novel KG-based interfaces and dashboards and focus on the ORKG resource comparison interface.

3.1 Recent Developments in Psychometric Analysis

In the literature on psychometric analysis, it was emphasized that a scale validation should be conducted as an iterative process that requires a multi-method assessment [220]. In other words, even when the results of a psychometric study are convincing and exhaustively reported, it is always beneficial to re-examine the main characteristics of the instrument with more rigorous, or newly developed, or simply different methods. Thus, the twofold goal is achieved: the scale is more comprehensively validated, and methodology of psychometric research is further developed.

An analytic procedure, or a psychometric protocol, for assessing item properties with CTT and IRT methods, was developed by Dima [58]. In terms of different ways of introducing method versatility, this procedure belongs to the consecutive use of these methods, when results obtained with one set of them is reassessed with the other. Psychometric analysis in accordance to this procedure consists of six steps that explore: (a) item descriptive statistics; (b) item properties according to non-parametric IRT, with homogeneity, monotonicity, local independence, and invariant ordering assumptions checked; (c) item properties according to parametric IRT requirements; (d) the structure of the scale according to EFA, CFA, VSS and ICLUST; (e) reliability of the scale and item properties according to CTT; (f) score statistics and distributions. To explore dimensionality of a scale, commonly used EFA and CFA can be combined with less

frequently employed ICLUST and VSS, and with still underused nonparametric IRT (MSA). Results of these methods can be compared to reach a conclusion about the structure of the scale. For reliability assessment, Cronbach's alpha with the confidence interval, McDonald's omega with the confidence interval, Guttman's lambda-6, and worst split half reliability (beta) are estimated. This allows for comparison between different indices and for correction of possible bias. To evaluate item functioning, methods of CTT and IRT are used, with the latter enriching the former with person fit measures and graphical presentations such as item trace lines and the person-item map. In terms of method versatility, this can be categorized as the consecutive use of different methods, as CTT, nonparametric IRT, and IRT (RSM) are used consecutively in the psychometric protocol.

In psychometric studies, novel advanced methods for determining convergent and discriminant validity of psychometric instruments were discussed. In terms of correlation analysis, the heterotrait-monotrait ratio of correlations criterion was shown to be an effective estimator of correlations between constructs [79]. In [210], different correlation-based techniques were compared, and recommendations on rigorous analysis were given. In addition to various forms of correlation analysis, more advanced methods were discussed in literature. In particular, SEM was recognized as an effective tool for demonstrating convergent and discriminant validity [220]. However, assumptions for SEM should be met in order for the analysis to be meaningful. Ignoring multivariate normality assumption was discussed as a major problem in studies using SEM [93]. Other methods to explore convergent and discriminant validity, as summarized in [22], include bivariate regression analysis, analysis of standard deviations of the differences between scores, and exploring the ICCs.

Hierarchical clustering of variables was suggested as a possible method for assessing the construct validity of psychometric instruments [45]. In hierarchical clustering, variables that are strongly related to each other, and thus contain similar information, are united in the same cluster. This clustering method can be either divisive or agglomerative. Agglomerative clustering starts with each variable forming a separate cluster, and the number of clusters is reduced at each stage based on a similarity or dissimilarity criterion, until all units are agglomerated in a single cluster. In divisive hierarchical clustering, the process goes in the opposite direction: all variables are initially united in a single cluster and then separated into different clusters. Various clustering algorithms and dissimilarity criteria can be used for this purpose; in [45], for instance, complete linkage method and Ward's linkage method are discussed and compared with FA. The ClustOfVar package in R implements an approach of agglomerative hierarchical clustering that can be used for various purposes. The authors defined a synthetic variable of a cluster as the first principal component in PCA [43]. The similarity (or homogeneity) criterion for numeric variables is calculated as squared correlations to the synthetic variable. Hierarchical clustering of variables as a method of construct validity analysis might be insufficient as it lacks mathematical precision such as in [210]: hierarchical clustering in its most

frequently used implementations does not account for the measurement error. It is possible to include the measurement error in clustering but such an algorithm would become excessively complicated. Even in its simplest form, hierarchical clustering is rather computationally expensive. Therefore, this approach is suitable for rather small (< 250) samples and as an additional means of assessing construct validity; specific values can be obtained with other methods.

Hierarchical clustering of variables might be helpful for presenting a general picture (a nomological network) with various levels of granularity. In Figure 3.1, an example of a dendrogram presenting the results of hierarchical clustering of variables is shown, with the names of the variables removed. The height of the dendrogram indicates the values of the aggregation criterion. With the chosen level of granularity, as depicted by the green horizontal line, three clusters are selected (these are indicated with different shades of blue). It is possible to choose a different height, or a different level of granularity, to obtain a partition of the data into four of more clusters. This approach can be useful when a researcher needs to explore a nomological network of a construct, so that the scores on the scale under scrutiny are compared with the scores on a number of theoretically related and unrelated scales.

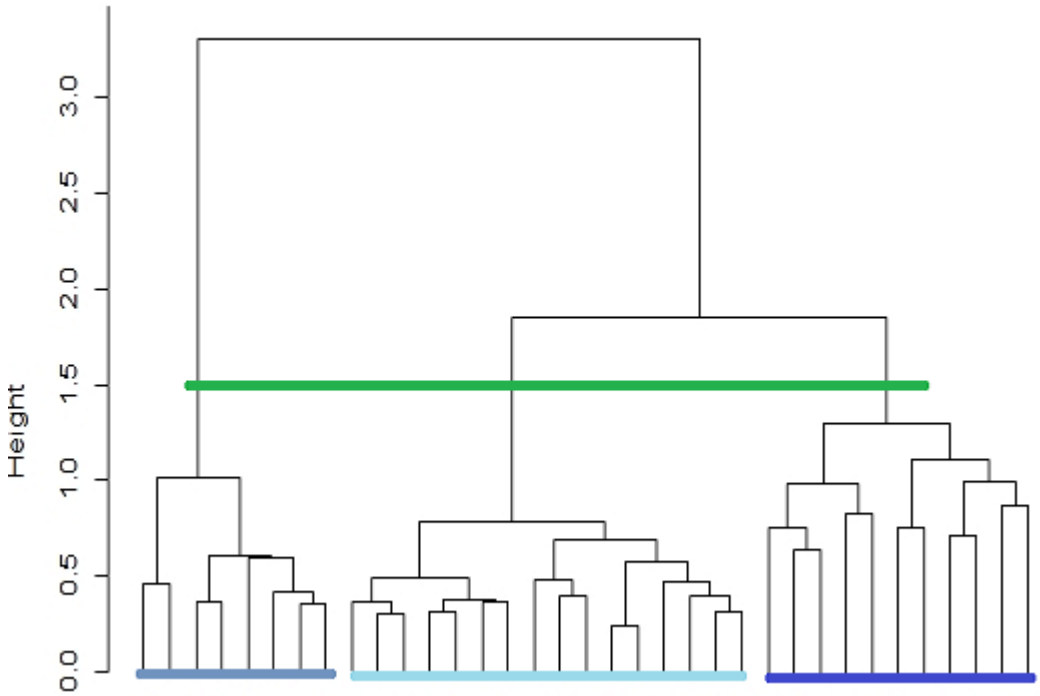

Figure 3.1: **Hierarchical Clustering of Variables.**

To summarise, a few developments in the area of psychometric analysis are relevant to my work on increasing method versatility in validation of psychometric instruments. These include (i) the psychometric protocol with CTT and IRT methods and (ii) hierarchical clustering of variables in application to validity analysis.

3.2 Recent Developments in Educational Data Analysis

In this section, I discuss recent developments in educational data analysis. These include statistical and ML techniques applied to large-scale educational surveys, such as PISA. As research on ICILS teacher data is scarce, and in the context of this work I am interested in cluster selection strategies, in the unsupervised learning subsection I focus on various suggestions for selecting the number of clusters in LCCA that can be useful for my analysis.

3.2.1 Machine Learning and Statistical Methods

A large amount of high-quality data provided by large-scale educational surveys can be analysed by means of various ML techniques [214]. On the NEPS website [169], multiple projects using ML methods are presented that aim to explore panel attrition, predict consequences of educational choices, and identify prospective university drop-outs. For occupational coding in NEPS and other surveys, naïve Bayes and Bayesian multinomial were used in [20]. For PISA 2012, different classification and dimension reduction algorithms, including RF and nearest neighbours, were tested on the Finnish data [214]. In [82], low rank matrix factorization was applied to the missing data imputation, and boosted regression trees models were trained for classification tasks on the Australian PISA data. Model agnostic methods, such as exploring feature importance and building partial dependency plots, were used in this study to make the results of the tree models more interpretable. In another study, RF was involved in feature selection on the process data in PISA 2012 [100]. Generally, RF was used in educational studies in different contexts. It was applied to predict the drop-out versus graduation results after the first two semesters [18], student success in an introductory statistics course [233], academic achievements based on students' cognitive psychological tests results [92], and the student progression in a virtual learning environment [102].

Multilevel statistical approaches to PISA, with school-level predictors included in the models, were widely used. With a two-level model for the Hong Kong data in PISA 2006, it was found that the school size is positively associated with students' science literacy scores [242]. A two-level HLM for the Indian data in PISA 2010 showed that school-level predictors influenced students' scores in reading, mathematics and science [10]. A multilevel mediation model was used on the Turkish data for the relationship between school-level ESCS and students' math anxiety, self-efficacy, and achievement in PISA 2003 [267]. In some cases, intraclass variance could be low; for instance, in a two-level HLM, school-level variables in the Irish data explained only 1% of variance in students' science achievement in PISA 2006 [49]. A multi-group exploratory SEM and a multivariate regression model was applied to PISA 2015 data (China and Germany) to explore relationships between students' attitudes to ICT and their mathematical, reading, and scientific literacy [162]. A three-level HLM involving 44 countries participating in PISA 2015 was applied to study relationships between various ICT factors and students' mathematical, reading, and scientific literacy [119].

Mutually enriching statistical and ML perspectives are integral part of data science [59]. In educational data analysis, integration of ML and statistical methods is yet to be obtained and has only recently started to be practiced in developing subfields of educational data mining and learning analytics (see [204] and [211]). However, statisticians and data analysts have discussed integration of ML and statistics from theoretical perspectives for rather long time [81]. Breiman [29] urged statisticians to embrace algorithmic thinking typical for ML. In this line of thought, novel methods were developed which combine ML and statistics. A combination of ML and statistical methods for training the most effective predictive model on the dataset in question was presented [105]. A bi-dimensional approach was suggested that takes into consideration explanatory power and predictive accuracy of a model, thus combining statistical and ML approaches to model evaluation [229]. Statistically reinforced ML was introduced as an approach, in which ML models are augmented by investigating significance and effect sizes typical for statistical models [213]. In terms of different ways of introducing method versatility, these belong to the simultaneous use of methods.

Another way to integrate ML and statistics is to select tools from either of these toolboxes depending on the study aims and the dataset characteristics. The best techniques for specific tasks can be chosen based on existing research on their comparative effectiveness. In a large-scale benchmarking experiment, RF was compared with logistic regression for binary classification tasks, and it performed better in terms of accuracy than this commonly used statistical method in 69% of the datasets [50]. In another study, eight ML methods (multi-layer perceptron, Bayesian neural network, radial basis functions, generalized regression neural networks also called kernel regression, k-nearest neighbour regression, CART regression trees, support vector regression, and Gaussian processes) compared unfavourably in terms of accuracy with statistical ones [156]. In [33], RF was compared in a simulation study with statistical inference used for the task of identifying genes associated with a phenotype; the authors showed that statistical and ML methods can be used complementary. Thus, a specific method for a specific task needs to be thoroughly selected. In this work, the toolbox approach to method versatility is applied to ML and statistical techniques in analysis of the PISA data.

3.2.2 Custer Selection in Model-Based Clustering

LCCA, or LCA [1], has developed rapidly to include ordinal data analysis [56], confirmatory methods [222], growth models [128], and Bayesian approaches [142]. In educational science, this method is increasingly applied to large-scale surveys data. Multilevel LCCA was used to explore the relationships between the curriculum and mathematical achievements and strategies in Dutch primary school students, and a considerable teacher effect was found [65]. LCCA was used to cluster students in

[1] As I explained in section 2.3.2, LCCA is called LCA when discussed from the statistical perspective rather than from the perspective of clustering. To avoid confusion, hereinafter in this work I call the method LCCA.

Taiwan based on their mathematical learning strategies as assessed by PISA 2012 [146].

In regard to ICILS data, students' responses to ICILS 2013 were explored by means of LCCA, with findings specifying the influence of extended or insufficient computer use on students' computer and information literacy [32]. In another study based on ICILS 2013 data, LCCA was applied to explore teachers' views on ICT in three European countries [63]. Both studies relied on model fit criteria, such as the BIC and the Akaike's Information Criterion (AIC), for cluster selection.

In the context of this work, I am primarily interested in different approaches to cluster selection in LCCA. Various information criteria for LCCA cluster selection were suggested and compared [170]. These studies gave mixed results, and there is no universally accepted criterion for choosing the number of clusters [234]. It was shown that the BIC is useful when the sample size is sufficiently large, and for small samples, the AIC is appropriate [61]. Although the BIC has a number of advantages over other information criteria [170], overreliance on the BIC as a single criterion was not recommended [61]. However, for a long time it was the only criterion implemented in commercial software, such as Latent Gold [195].

In other studies, criteria used for selecting the number of clusters in LCCA were not restricted by model fit indices. For instance, it was suggested that the elbow heuristic for the BIC plot can be applied for this purpose [74], [171]. This means that the researcher determines the "elbow" of the plot, after which the change in successive values becomes less noticeable. The elbow heuristic implies the subjectivity of choice but is effective and simple, and therefore is common in EFA [116] and in cluster analysis [23].

The stability of cluster partitions was shown to be an important criterion for selecting the number of clusters. An estimation scheme for clustering instability was developed [68] to inform the selection of the number of clusters so that the corresponding estimated clustering instability could be minimized. In Hennig's works on clustering (see, e.g., [110]), different methods for the stability assessment were suggested, including splitting the dataset or "jittering" the observations; the bootstrap validation procedure was shown to be an effective method of cluster selection.

Moreover, Hennig and Liao maintained that a criterion typically used in distance-based clustering, the ASW, which is needed for checking whether clusters have relatively small within-cluster dissimilarities, could be an aid in model-based clustering such as LCCA [111]. It is not a common practice to use distance-based criteria for model-based clustering, and this approach was rather unorthodox. After that, Anderlucci and Hennig [7] applied the ASW criterion to LCCA models and showed that LCCA can perform at least as well in terms of the ASW as distance-based methods. Thus, the long established routine of relying solely on model fit indices, or even the single fit index (the BIC), was supplemented with alternative criteria for selecting the number of clusters in LCCA.

To summarise, a few developments in supervised and unsupervised learning methods applied to educational data analysis are relevant to my work. These include (i) previous studies applying HLM, RF, or LCCA as a clustering method, to PISA and ICILS data; (ii) works that integrate, or combine in the frame of the toolbox approach, statistical and ML methods; and (iii) studies on selecting the number of clusters in LCCA based on a different set of criteria than solely fit indices.

3.3 Recent Developments in Knowledge Graph-Based Interfaces

Novel user-friendly KG-based interfaces are continuously created that aim at facilitating acceptance of this technology by wider groups of users, including non-experts, and various aspects of technology acceptance related to KG-based interfaces are explored. In order to reduce cognitive load of the user, a visualised knowledge map for semantic tagging was suggested, and its effectiveness was evaluated with a field experiment [120]. Development of an interface with easily understandable visualisations that allows both experts and novices querying a knowledge base was described in [136]. The user perception of two spatial presentations of web search interface, hierarchical tree and graph view was studied in [217]; the findings confirmed that the hierarchical tree interface can be more effectively employed for getting an overview of a field, and the graph interface for answering specific questions. A visual interface for non-experts querying KGs was developed, and a usability evaluation showed that it compares beneficially with the baseline system used by Wikidata [247]. A linked data-powered framework for scholarly information management was launched at the state level in Ecuador; the core component of the system was user-friendly graphical user interface [187].

Dashboards as dynamic visualisations are increasingly used in different areas of research and practice. These include, for instance, dashboards presenting the epidemiological information to the general public [40]. In [67], a dashboard was described that visualised patients' information for hospital intensive care units. The dashboard aimed to reduce cognitive strain experienced by the clinicians, and it was shown that these intuitive visualizations were useful for rapid information assimilation and pattern recognition facilitating diagnostic insights. In the context of urban mobility, KG-powered dashboards are also applicable, as they resemble instruments that humans are accustomed to using regularly. For instance, in [216] development of an application was reported that performed metadata analysis to automatically generate dashboards displaying various mobility indicators. In the area of academic interaction, a dashboard presenting information about scientific conferences was designed for less technically-savvy audiences; it integrated statistical analysis, semantic technologies and visual analytics, so that the user could visualise several metrics of a specific conference [9]. In [143], the science citation knowledge extractor is described, which employs natural language processing and machine learning to retrieve key information from scientific publications and present interactive data visualizations. With this tool, biological and

biomedical researchers can understand how their work is being utilized by the academic community.

In the frame of the ORKG research service infrastructure initiative, the ORKG resource comparison was implemented as an SKG-based interface [174]. A research contribution includes a research problem addressed by an academic publication, a method (or methods) it uses and the research results. The resource comparison interface allows selecting research contributions, mapping their properties (such as methods, results, and others), and publishing the resulting comparison online in a tabular form.

Figure 3.2: **The ORKG Resource Comparison.**

In Figure 3.2, I show an example of a resource contribution comparison on the topic of attitudes towards ICT in PISA 2015 and 2018. The presentation is customizable: the user can enable or disable properties to be shown in the tabular form. Thus, relevant publications can be selected that, for instance, apply specific methods to the topic of interest. The comparison can be shared and exported in different formats. According to the evaluation of user performance, the participants found the service useful and fairly intuitive [174].

To summarise, related work in the area includes (i) KG-based interfaces and (ii) various dashboards, which are examples of multi-relational dynamic visualizations. In my work on extending the range of means for scholarly communication, I relied on the principles used in these studies to create a dashboard that could be complementary to the ORKG resource comparison, an effective and intuitive interface implemented in the frame of the ORKG research service infrastructure initiative.

3.4 Summary

In this chapter, I outlined previous studies related to my research questions and contributions to the thesis. The literature surveyed was relevant to this thesis either in terms of methodology, or in terms of findings that could be used in my research.

The first section described recent developments in psychometric analysis, on which I can base my own work. The psychometric protocol combining CTT and IRT methods to explore dimensionality, reliability, and item functioning of a scale can be a useful aid in facilitating method versatility in psychometric analysis. As this protocol does not include procedures for assessing construct validity, I referred to hierarchical clustering of variables used for this purpose. There are various implementations of hierarchical clustering, and I outlined a specific PCA-based implementation that I selected for this purpose.

In the second section, I reviewed literature on supervised and unsupervised learning in data analysis of large-scale educational surveys, such as PISA and ICILS. In regard to supervised learning, I outlined applications of ML methods, such as RF, and multilevel methods from the statistical toolbox to educational data analysis and specifically to PISA. After that, I discussed previous works on various combinations of ML and statistics, which is yet to be accepted in educational data analysis. In regard to unsupervised learning, I focused on cluster selection in LCCA and discussed the studies showing that strategies based solely on model fit were suboptimal, and more versatile selection procedures would be beneficial. Resorting to the distance-based ASW criterion was suggested in some works, while others relied on the BIC elbow heuristic or the bootstrap validation procedure.

In the third section, I discussed related work on KG-based interfaces, including the ORKG resource comparison interface, and on dashboards from various domains. Thus, my work on increasing method versatility in data analysis of human attitudes towards technology at stages of measurement, research, and communication, is included into the context of related studies.

Chapter 4

Multi-Method Approach to Validating a Scale

In Chapter 2 (section 2.2), I discussed the background of psychometric research, and in Chapter 3 (section 3.2), outlined recent developments in the area. This chapter presents my work on method versatility in psychometric analysis, which I conducted to deal with RQ1.

> **RQ1.** How to facilitate method versatility in validation of psychometric instruments related to human attitudes towards technology?

As I explained in Chapter 2, various ways of introducing method versatility can be found in literature. I chose the consecutive use of different methods for the psychometric work, as reassessing a scale developed by other researchers was consistent with the idea of iterative multimethod assessment [220]. In order to get the data for reassessment, I selected a recently developed scale that (i) intended to measure a construct related to human attitudes towards technology, (ii) could be considered useful from the domain knowledge perspective, and (iii) was validated with commonly used, relevant, and transparently reported methods. This was the ATI scale by Franke, Attig, and Wessel [80], and when I contacted the authors, they were willing to share the dataset. Results of their study showed that the ATI scale was a valid unidimensional instrument with high reliability. Their methods included CTT assessment of item functioning and EFA for dimensionality, with parallel analysis as a factor retention method and PFA as a factor analytic method. For reliability, the authors reported the value of Cronbach's alpha, and construct validity of the scale was indicated by Cohen's correlations with other constructs. These methods were sufficient according to the standards of the field, in which they were - and still are - the most commonly used methods. They were reported transparently and gave space for reassessment with the aim of reconfirming the results of the analysis and increasing method versatility in the area.

4.1 Analytical Strategy

The analytical strategy included applying nonparametric IRT (MSA) and parametric IRT (RSM) methods to reassess the results obtained with CTT methods on dimensionality, reliability, and item functioning of the ATI scale. Hierarchical clustering of variables was used to reassess the results for convergent and discriminant validity obtained by the authors with correlation analysis. In terms of method versatility, it was the consecutive use of different methods for the same task of psychometric assessment, as depicted in Figure 4.1. The data analysis was conducted with R, version 3.5.2 [196]. A coherent system of packages tidyverse was used for data manipulation and visualization [260]. The R script is available on GitHub https://github.com/OlgaLezhnina/ExtPsych-ClustATI.

Exploring and imputing missing data was not requited, as there was no missingness in the dataset. Prior to the analysis, the decision on outlier removal was made. As there were no valid reasons to remove the outliers (see, for instance, [15]), the cases with aberrant response patterns (Guttman errors) and multivariate outliers were kept in the dataset for further analysis.

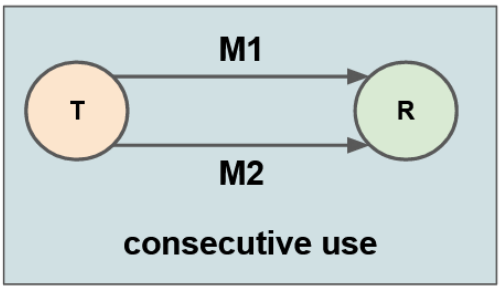

Figure 4.1: **Method Versatility: Consecutive Use.**
T stands for task, M for method, R for result.

For convergent and discriminant validity assessment, I used hierarchical clustering of variables. I did not repeat the correlation analysis conducted by the authors, as the latter was reported in [80] and could be easily reproduced by an interested researcher. For agglomerative clustering, I selected the ClustOfVar package [43]. The homogeneity criterion of a cluster was defined as the sum of correlation ratios to the synthetic variable (the first component obtained by PCA). The stability of partitions was evaluated with the bootstrap approach, and the default of 100 bootstrap samples was used. Dendrograms were built to visualize the results. The analysis was conducted both (i) on the mean values of all scales included in validity analysis (see section 4.2), and (ii) on all items of these scales. The former was needed to reassess the findings obtained by the ATI authors, as their correlation analysis was also conducted on the mean values of the scrutinized scales. The latter allowed a more detailed picture of the nomological network of the ATI scale to be obtained.

4.1. Analytical Strategy

Dimensionality, reliability, and item functioning were assessed by the authors of the ATI scale with CTT methods and reported in [80]. Some of the procedures, however, were repeated in the current analysis, as it was conducted in accordance with [58]. This psychometric protocol includes assessment of (i) item descriptive statistics; (ii) item properties according to nonparametric IRT; (iii) item properties according to parametric IRT; (iv) dimensionality with CTT methods; (v) reliability and item properties with CTT methods; (vi) score statistics and distributions. Thus, nonparametric IRT, parametric IRT, and CTT methods were applied in blocks that could be useful for fitting the models in accordance with each of these frameworks. Here, however, the analytical strategy and the results of assessment are described in a different order, for logical consistency: (i) dimensionality analysis, (ii) reliability assessment, and (iii) item analysis, as assessed by CTT and IRT methods. I report analytical choices not covered by the protocol, as well as my amendments to the procedures that constitute it, with references to sources that gave me grounds for these decisions.

In the frame of dimensionality assessment, CFA, EFA, VSS, and ICLUST were used as traditional CTT methods. CFA was conducted with the MLR estimator. The following rules were applied to assess the global fit between the tested model and the data [117]: the chi square test should reveal no significant differences between the model and the observed covariances ($p \geq .05$); the thresholds for fit indices should be CFI $\geq .95$, RMSEA $\leq .08$, and SRMR $\leq .08$. I made a few amendments to the procedure described in [58] regarding the assumption checking: multivariate normality of the data was tested with Mardia's test of multivariate normality [93], and the package QuantPsyc [73] was selected for the implementation of the test, as it was proved to give unbiased results [127]. In addition, I explored modification indices of the CFA model [93]. In EFA, PAF was used as a factor analytic method. For factor retention, scree plot analysis, parallel analysis, and the acceleration factor (which is a numeric expression of the scree plot inspection results) were used. As my amendments to the procedure, the KMO and the Bartlett's test of sphericity were conducted with the psych package [202], and for the factor extraction, the nFactors package was used [198]. These decisions were informed by existing guidelines for EFA [71], [116]. VSS was conducted, and the Velicer MAP results obtained. Also, the hierarchical clustering of the items was conducted, and the results of ICLUST with alpha and beta values for the subscales were presented graphically.

Dimensionality of the ATI scale was reassessed by methods of nonparametric IRT (MSA). Homogeneity values for the scale and for each item with standard errors (SE) were obtained. An Automated Item Selection Procedure (AISP) was conducted to explore scalability of items and dimensionality of the scale at increasing threshold levels of homogeneity from .05 to .80. Assumptions for MSA (monotonicity, local independence, invariant item ordering) were checked. The monotonicity test and the invariant item ordering test were conducted with the default minisize (n = 80). The threshold of the Crit value = 40 for the monotonicity test was taken; ideally, an item should have the Crit value of 0, as recommended in [225].

Internal consistency reliability of the ATI scale was assessed with alpha, beta, lambda-6, and omega estimates. Confidence intervals for alpha and omega coefficients were reported.

Item functioning was assessed in the frame of the CTT approach, as to include frequencies of endorsement, the inter-item correlation matrix, item-total associations, the distribution of total scores, item discriminations (corrected item-total correlations) and Cronbach's alphas when the item is removed. A detailed picture of item functioning was obtained by parametric IRT methods. RSM was fit to determine whether the ATI scale satisfies requirements for additive measurement. Item fit was explored; mean square values from 0.6 to 1.4 were taken as acceptable, while standardized fit statistics values above 2 were considered not suitable for measuring the latent construct on an interval level [58]. Item trace lines were built to explore the relationship between the latent trait and the probability to endorse a specific category of the item. Person fit was evaluated with the same thresholds as item fit. Separation reliability of the scale (with the cut off value of 0.80) and person separation (with the cut off value of 2) were assessed. The person-item map was built to visually explore how the items relate to the latent trait continuum.

In the process of dimensionality testing, one of the items appeared to be less scalable than the others. Further analysis was conducted on (i) the initial scale and (ii) the scale with the less scalable item excluded. The results for both versions were presented, as they give information about the scale and the item relevant to the specific sample; further research on a much larger and diverse sample is required to make decisions about any possible changes in the scale.

4.2 Data

The dataset (N = 240) was shared by the authors of the ATI scale on my request (and is currently in free access at https://ati-scale.org/). The data was collected by means of the MTurk in the USA. Demographic variables were not included in the current analysis. There were scores on 12 scales in the dataset: (a) the ATI scale, nine items; (b) Technical Problem Solving Success (TPSS, four items); (c) Technical System Learning Success (TSLS, three items); (d) Interest in Technology (Interest, four items); (e) Need for Cognition (NFC, four items); (f) Geekism (GEX, 15 items); (g) a short form of Big Five Inventory (BFI-10, five scales, each with two items); and (h) Regulatory Focus Scale adapted for technical systems (RFC, six items). Reversely coded items were recoded. Hereinafter, the reversely coded items (in the ATI scale, these are ati03R, ati06R and ati08R) are indicated with the letter R.

There were no missing data in the scale dataset due to strict quality filtering, including completeness check. As the authors of the scale reported, those respondents who (a) did not complete their survey, (b) completed the survey twice, (c) failed to answer the built-in attention checks, or (d) resided outside the USA were excluded

from the dataset. For more information on the sample, the scales, and the quality filtering, see [80].

4.3 Results

Convergent and discriminant validity of the ATI scale was explored with hierarchical clustering of variables. The analysis was conducted for the means of the 12 scales involved, and the stability of partitions was checked with the default of 100 bootstrap samples. The results suggested that the eight-cluster partition and the nine-cluster partition were most stable. The eight-cluster partition was explored: the ATI construct formed one cluster with Geekism, Interest in Technology and RFS constructs. In case of the nine-cluster partition, the ATI formed one cluster with Geekism and Interest in Technology constructs. For the most parsimonious two-cluster partition, the results were similar (see Figure 4.2).

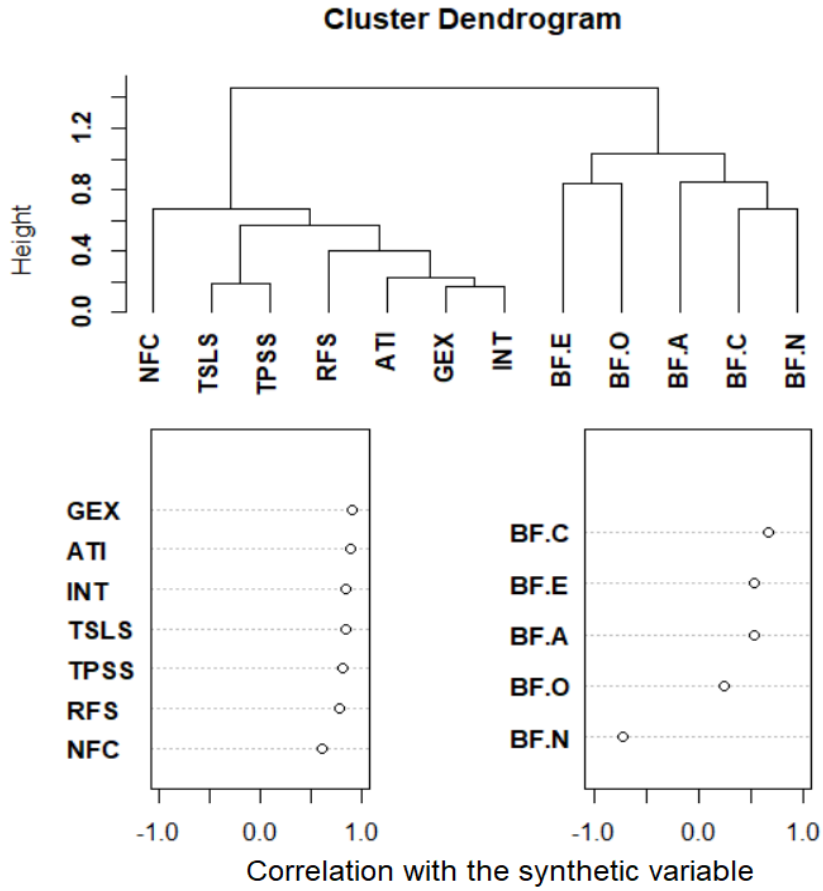

Figure 4.2: **Dendrogram and Two-Cluster Partition of the Means.**
The scales are labelled with abbreviations used in [80]. NFC is Need for Cognition; TSLS is Technical System Learning Success; TPSS is Technical Problem Solving Success; RFS is Regulatory Focus Scale adapted for technical systems; GEX is Geekism; INT is Interest in

Technology; BF are Big Five scales: BF.E is Extraversion; BF.O is Openness; BF.A is Agreeableness; BF.C is Conscientiousness; and BF.N is Neuroticism.

The dendrogram for the two-cluster partition, with the height indicating the values of the aggregation criterion, is presented in the upper part of Figure 4.2. The lower part of the figure depicts correlations of cluster elements with the synthetic variable in both clusters. The ATI was closely related to such constructs as Geekism and Interest in Technology, and other technology-related constructs were in the same cluster with it, while the Big Five dimensions (Extraversion, Openness, Agreeableness, Conscientiousness, and Neuroticism) formed another cluster. The gain in cohesion for the two-cluster partition was 20.59%, for the eight-cluster partition 85.96%, and for the nine-cluster partition 91.60%.

Hierarchy of variables was also constructed for all items of all scales. The ATI items were close to Geekism, Interest in Technology and Need for Cognition items. The stability of partitions was explored with the default of 100 bootstrap samples. The two-cluster partition showed the relatively high stability according to the ARI, with the next stability maximum at 26 clusters. The two-cluster solution was explored: it gave a gain in cohesion of 6.99%, and all items of the ATI scale were close to items of other technology-related scales in one of the clusters, with Big Five items in the other cluster. The 26-cluster partition was explored. The gain in cohesion was 71.19%. Items of the ATI scale formed clusters with items of Geekism, Interest in Technology, and Need for Cognition scales.

Overall, the results of hierarchical cluster analysis confirmed the findings by Franke, Attig, and Wessel [80], who reported that the scale had high correlations with technology-related constructs and low correlations with the Big Five constructs. However, the method that I used to reassess construct validity, hierarchical clustering of variables, allowed for more granular picture of the nomological network: for any cluster partition, it could be seen how the constructs relate to each other. The same is the case for all items of the scrutinized scales (see the dendrogram in Figure 4.3).

Validity analysis required other scales to be involved; the remaining procedures of psychometric assessment were conducted only on the items of the ATI scale. In my reassessment of dimensionality, reliability, and item functioning of the ATI scale, I consecutively (i) used the same procedures and criteria as the authors of the scale (EFA, Cronbach's alpha, CTT methods), and (ii) reassessed the results with parametric IRT and nonparametric IRT methods.

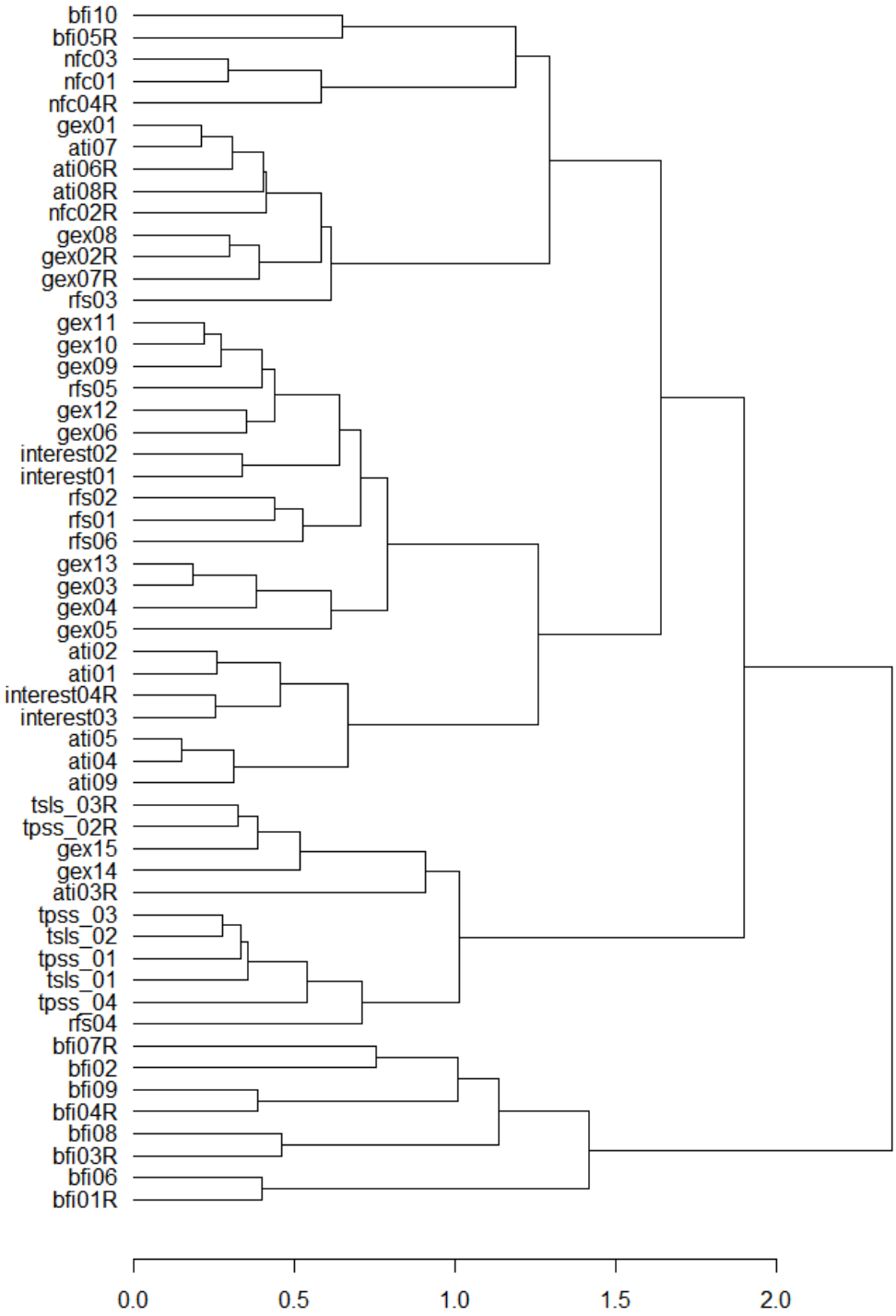

Figure 4.3: **Dendrogram of the Items.**
The scales are labelled with abbreviations used in [80]; 'R' means that the item is reversely coded.

Prior to other procedures, multivariate normality of the data was studied with the Mardia's test as my amendment to the protocol, as the latter includes checking multivariate outliers based on the Mahalanobis D2 values but not on the Mardia's test. The results of the test for skew and kurtosis were significant, with $p < .001$, which meant that the multivariate normality assumption was violated. Aberrant response patterns of respondents were explored with analysis of Guttman errors. There were 13 outliers (cases with a number of Guttman errors higher than the cut-off value of 65.5). As there was no valid reason to remove the outliers, the cases were kept in the dataset for further analysis.

For dimensionality analysis, CFA with the MLR estimator was conducted on the ATI scale, with one factor as known from [80]. The results did not show a good fit of the model. The chi square test was significant $\chi 2\ (27) = 231.34$, $p < .001$. As for the fit indices, the values of CFI = 0.84, RMSE = 0.19, SRMR = 0.07 indicated insufficient fit. CFA for the ATI8 scale also did not show good fit, and changes based on specification search with modification indices were not able to significantly improve it. The situation can be explained by the fact that the data was not multivariate normal, and CFA, even with a robust estimator, is generally not recommendable in this case.

EFA was conducted on the ATI scale. The KMO verified the sampling adequacy for the analysis. For the scale, the KMO was .89, and KMO values for individual items were above .85. Bartlett's test of sphericity, $\chi 2\ (36) = 1448.93$, was significant with $p < .001$, thus assumptions for the EFA were met. PAF was used as a factor analytic method. To determine the number of factors, parallel analysis, scree plot analysis and the MAP test were used. All methods supported one-factor solution, as shown in Figure 4.4.

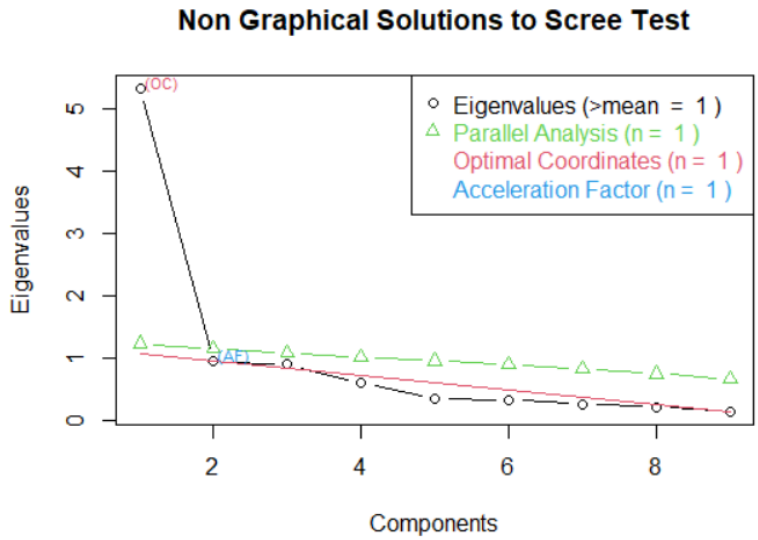

Figure 4.4: **The EFA Scree Plot.**

The decision to retain one factor was supported by VSS analysis, which indicated that the first level of complexity achieves a maximum of .92 with one factor. MAP achieved a minimum of .06 with one factor. ICLUST also gave one cluster solution, which is graphically presented in Figure 4.5.

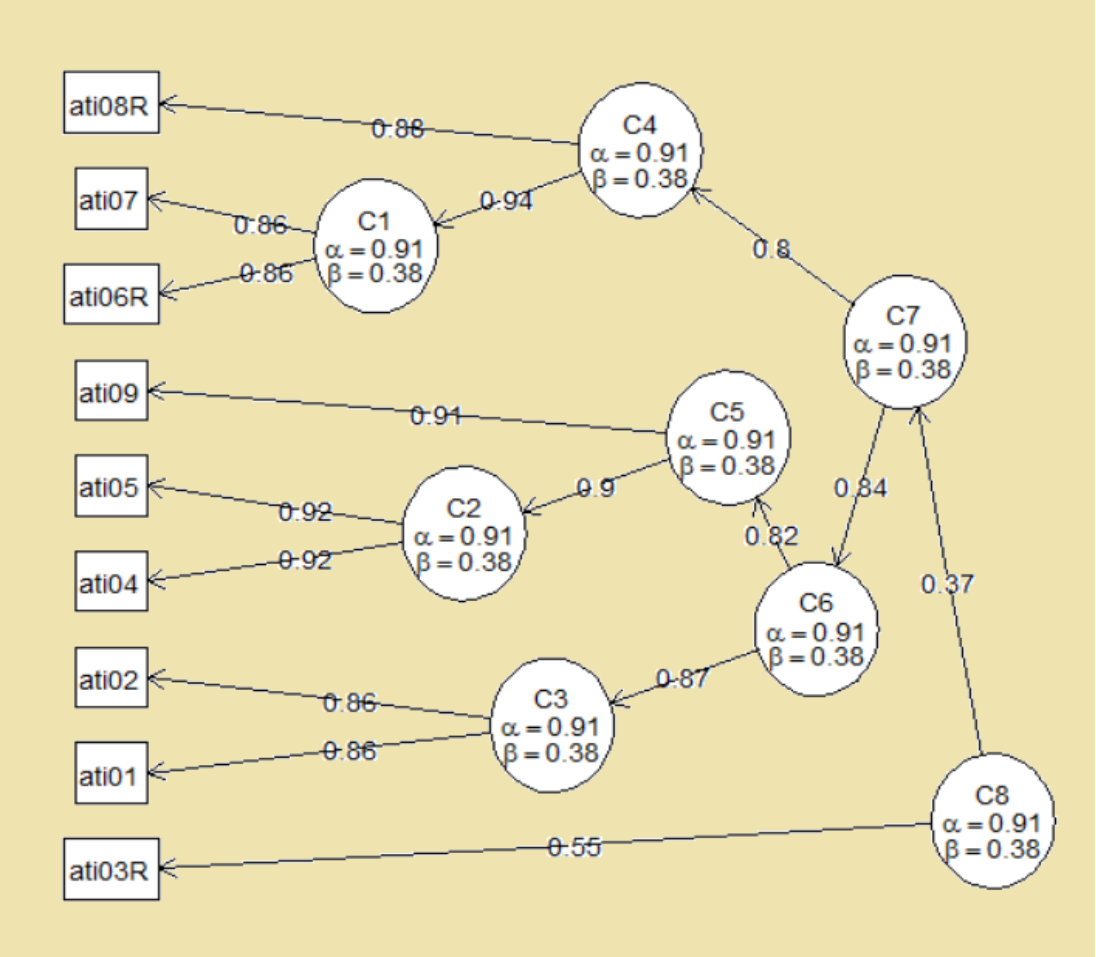

Figure 4.5: **The ICLUST Visualization.**

Therefore, one factor, which explained 55% of variance, was retained in the final analysis. Standardized factor loadings for all items except one ranged from .68 to .85, while for item ati03R the loading was .28.

Dimensionality of the ATI scale was reassessed with nonparametric IRT (MSA) methods. An AISP was conducted to explore scalability of items and dimensionality of the scale at increasing threshold levels of homogeneity. According to the AISP results, the ATI scale is unidimensional. As the minimum threshold level for homogeneity is .30, items with value 0 at this level or below are considered unscalable. For the ATI, there was one item (ati03R), which showed a lack of scalability at the threshold as low as .25 (see Table 4.1). Thus, it could be recommended to remove this item from the scale.

Further analysis was conducted for both versions of the ATI scale, the current version (ATI) and the eight-item version with item ati03R removed (hereinafter called ATI8). The results are reported separately whenever the comparison between the two versions is meaningful; in other cases, the results for the original version (ATI) are reported.

Table 4.1: **AISP with Increasing Homogeneity Thresholds.**

Items	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80
ati01	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	0
ati02	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	0
ati03	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
ati04	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ati05	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ati06	1	1	1	1	1	1	1	1	1	1	1	1	2	3	3	0
ati07	1	1	1	1	1	1	1	1	1	1	1	1	1	3	3	0
ati08	1	1	1	1	1	1	1	1	1	1	1	0	2	3	0	0
ati09	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0

The complete item set of the ATI scale had a homogeneity value H = .55 with a standard error of .03, and the complete item set of the ATI8 scale had a homogeneity value H = .64 with a standard error of .03. Thus, removing item ati03R would lead to increase in homogeneity of the whole scale and to increase in homogeneity of all items, as can be seen from Table 4.2.

Table 4.2: **Homogeneity Values for ATI and ATI8 Items.**

Scales		ati01	ati02	ati03R	ati04	ati05	ati06R	ati07	ati08R	ati09
ATI	*H*	.57	.61	.23	.58	.63	.57	.60	.54	.60
ATI8	*H*	.62	.67	-	.64	.69	.63	.68	.59	.65

According to the local independence (conditional associations) test, all nine items of the ATI scale meet the local independence criterion. The monotonicity test (with the default minisize of n = 80) gave criterion values (Crit) of 0 for all items, except for item ati03R; for this item, the Crit value was 41. As the threshold for the Crit value is 40, and ideally, an item should have the Crit value of 0, the monotonicity test for item ati03R showed a violation of the assumption, while other items showed very good monotonicity. The invariant item ordering (IIO) test for the ATI scale (with the default minisize) showed that there were significant violations of invariant ordering for items

ati01, ati06R and ati07 (one violation per each item) and three significant violations of invariant ordering for item ati03R. The output of the test explicitly suggested removing item ati03R from the scale. When IIO test was conducted for ATI8 scale, it showed zero violations of invariant ordering for each item.

Results of reliability analysis the ATI and ATI8 scales are presented in Table 4.3. They include Cronbach's alphas with confidence intervals, McDonald's omegas with confidence intervals, Guttman's lambdas-6 and the worst split halt reliabilities (betas). Both scales, the original ATI and the ATI8, had excellent reliability according to all indices.

Table 4.3: **Reliability Indices for ATI and ATI8.**

Scale Version	Alpha	Guttman's λ-6	Beta	Omega
ATI	.90[.88-.92]	.92	.83	.90[.88-.92]
ATI8	.92[.91-.94]	.93	.86	.92[.89-.94]

Item functioning was explored with CTT methods. The ATI items showed sufficient variation to differentiate respondents on their affinity for technology interaction. Frequencies of endorsement showed that all response options were represented in the data (see Figure 4.6).

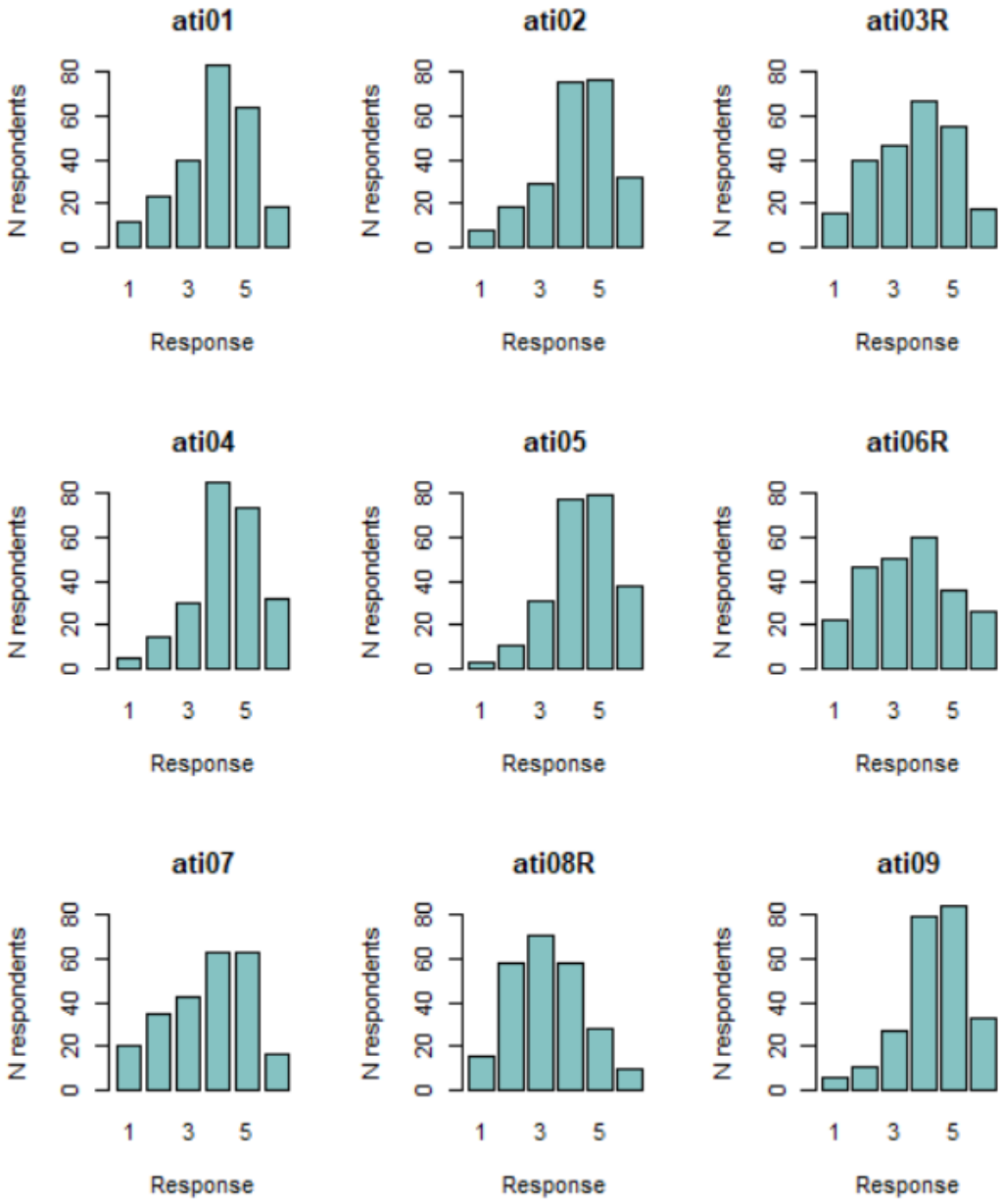

Figure 4.6: **Barplots for the ATI Items.**

Associations between the items were positive. Item ati03R showed the weakest correlations with other items, ranging from .14 to .26, while other items correlated with each other in range from .46 to .85.

Item discriminations (corrected item-total correlations) and Cronbach's alphas when the item is removed are presented for all items of the scale in Table 4.4. It can be seen that item ati03R has the lowest discrimination value, and its removal would lead to increase in the value of alpha for the scale.

Table 4.4: **Item Analysis (CTT).**

Items	ati01	ati02	ati03R	ati04	ati05	ati06R	ati07	ati08R	ati09
Item-total	.71	.77	.27	.73	.79	.71	.76	.66	.74
Alpha-rm	.89	.88	.92	.89	.88	.89	.88	.89	.89

"Item-total" stands for item-total correlation, "alpha-rm" stands for alpha when the item is removed.

In the frame of IRT, summary item fit for the ATI scale was explored. Criteria for item fit are the mean squares ranging from 0.6 to 1.4, and values above 2 are considered not suitable for measuring the latent construct on an interval level. Outfit and infit mean squares values of all items, except for item ati03R, ranged from 0.56 to 0.95. For item ati03R, the outfit value was 2.52, and the infit value 2.10, which was above the threshold. Person fit was evaluated based on the same criteria as item fit. There were no respondents with misfit according to outfit values or infit values. Item trace lines for all items, including ati03R (see Figure 4.7), showed sufficient discrimination for the response categories.

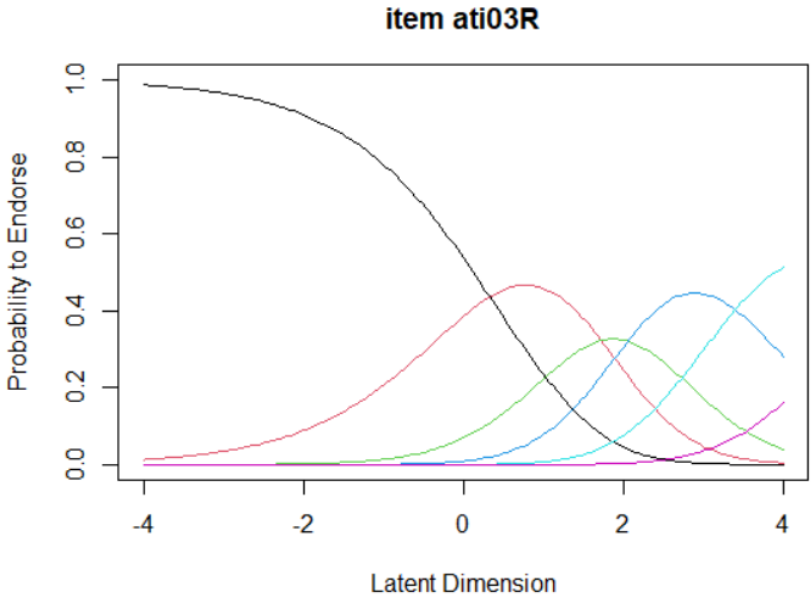

Figure 4.7: **Item Trace Lines for Item ati03R.**

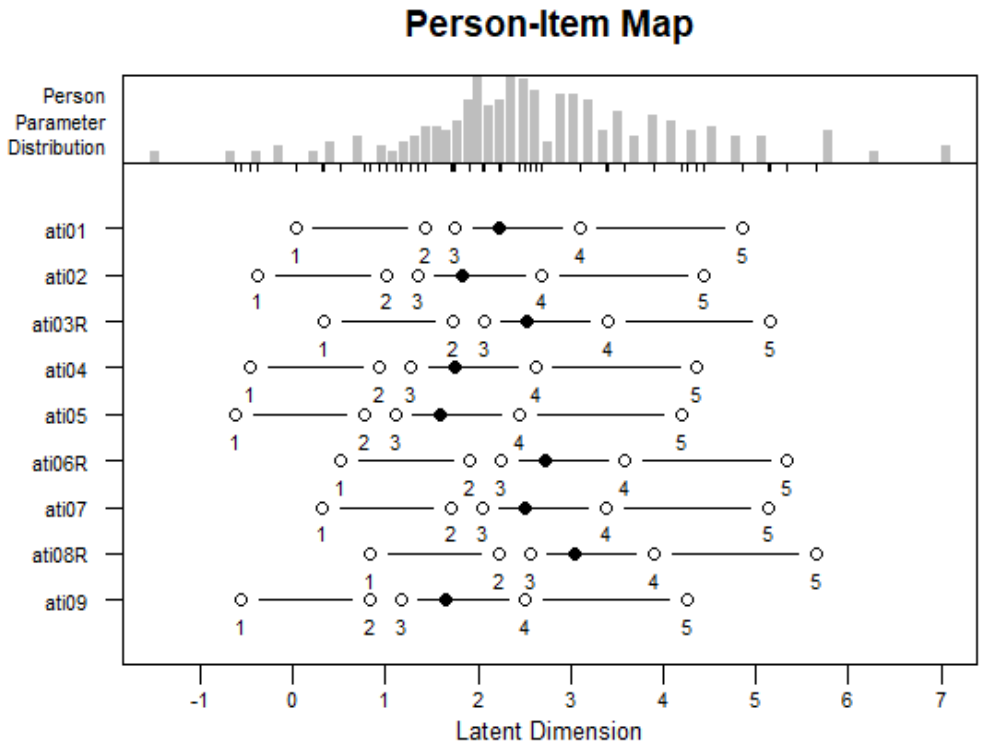

Figure 4.8: **Person-Item Map.**

The hierarchy of item difficulty and the match between person ability and item difficulty (scale targeting) were explored graphically via the person-item map (Figure 4.8). The upper panel shows the person parameter distributed along the scale. The lower panel shows item difficulties and thresholds. The circles represent thresholds for items, and bullets represent item difficulties. Neither disordered thresholds, nor redundant items can be detected. Items are located along the scale and thus able to measure various levels of the latent trait.

4.4 Summary

In this chapter, I showed how method versatility can be introduced at the measurement stage of the research cycle. I applied multi-method psychometric procedures to re-assess validity, dimensionality, and reliability of ATI scale. Method versatility was introduced as the consecutive use of different methods for the same task, that is, psychometric analysis of the scale.

In terms of the scale validation results, it was confirmed that the ATI scale is a valid and reliable instrument. My findings supported conclusions by the authors of the scale conceptualizing the Affinity for Technology Interaction construct as close to Geekism and Interest in Technology constructs and distinct from Big Five constructs. The results for less scalable item ati03R were reported; however, item characteristics obtained for a

specific sample should not be automatically generalised to the population [60]. Therefore, based on the results of extended procedures, it can be recommended to further explore the item functioning on a larger sample, and only after that, final conclusions can be made. Overall, the scale had good homogeneity and good ability to differentiate respondents on the measured construct, and it can be recommended for research of the area of human-technology interaction.

In methodological terms, this work exemplifies conducting psychometric validation as iterative multi-method assessment. I used hierarchical clustering of variables for construct validity analysis that commonly relies on correlations of constructs. It might be recommendable to explore the potential of this method in more detail, as it allows for various levels of granularity in relationships of different constructs, and thus, convergent and discriminant validity, or nomological network, of a scale can be assessed in detail. The dendrogram of all items of the twelve scales illustrates the benefit of this approach; exploring correlations of these items with each other would have been a cumbersome process. Methods of CTT and IRT were used to explore dimensionality, reliability, and item functioning of the ATI scale. In the dataset under consideration, the data was not multivariate normal; this case is not infrequent in the field [223], and neglecting this assumption would bias the results of CFA [93]. For dimensionality, I used nonparametric IRT (MSA) to reassess the findings obtained with common CTT methods. MSA does not impose the multivariate normality assumption on the data, and it gave detailed and concise results regarding the scale. Internal consistency reliability was reassessed, in addition to alpha, with beta, lambda-6, and omega indices. For item functioning, I used parametric IRT methods and applied RSM to reconfirm the results obtained with CTT. Person parameters, which are not included in the frame of CTT, are useful, as they take into consideration the sample-specific nature of the findings. The methods described in this chapter and the related R code can be used by researchers in the integral process of psychometric assessment of a scale or as separate blocks related to specific psychometric procedures.

Chapter 5

Combining Statistics and Machine Learning for Educational Data Analysis

In the previous chapter, I reported the results of my work on method versatility at the measurement stage of the research cycle and outlined the consecutive use of different methods for reassessing a psychometric instrument. In Chapters 5 and 6, I deal with the stage of research *per se* and address RQ2.

RQ2. How to facilitate method versatility in educational research on human attitudes towards technology?

This chapter relates to versatility in supervised learning methods, and Chapter 6 to unsupervised learning, used for data analysis of large-scale educational surveys. In the current chapter, I apply statistical and ML methods to educational data analysis on PISA, a survey that provides researchers with a large amount of thoroughly collected data. PISA is conducted every three years to measure 15-year-old students' literacy in different domains and other attitudinal and demographic factors. Literacy scores in PISA are translated into proficiency levels, from Level 1b to Level 6, using cut-off points [182]; students with scores below the baseline Level 2 are classified as low performers and students with scores at Level 5 and above as high performers [180], [184], [193]. Therefore, both classification tasks and regression tasks with the PISA data are feasible. For a classification task, a performance class can be predicted, or a performance level, or a binary task can be formulated. For a regression task, the outcome variable can be a student's score in a subject, such as mathematics, reading, or science.

Students' attitudes towards ICT (ICT competence, ICT interest, ICT autonomy, and ICT in social interaction) were selected as input variables because this thesis is focused on human attitudes towards technology. Taking ICT literacy as an outcome variable was not possible, as in contrast to some other surveys (such as ICILS), PISA does not measure actual performance in ICT. However, a vast amount of studies confirmed the

relationships between students' attitudes towards ICT and their scores in mathematics and science in PISA. Already in PISA 2003-2012, positive attitudes towards ICT were shown to be significantly and positively associated with mathematical and scientific literacy scores [98], [153], [191]. Recent findings on attitudes towards ICT in PISA 2015 showed that ICT autonomy was significantly positively associated with mathematical and scientific literacy scores in all countries participating in the optional questionnaire, and ICT in social interaction was significantly negatively associated with these scores in all countries [119]. For ICT competence and ICT interest, the significance and the sign of their relationships with literacy scores varied at the country-specific level [162], [173]. I used two recent PISA datasets, the 2015 and the 2018 data, and selected solely German samples to avoid pitfalls outlined by critics of PISA with regard to international comparisons based on simplistic interpretations [8], [88], [114], [271]. In order to determine the relative importance of attitudes towards ICT, I compared them with each other and with two demographic variables, which were shown to be influential factors for academic achievement in mathematics and science in Germany: ESCS and gender [179], [185], [208].

From the methodological perspective, my aim was to present a versatile way to apply statistical and ML methods to a large-scale educational survey. The most suitable tools from these two toolboxes were selected for each of the three consecutive tasks. For the first task, missing data imputation, I chose RF which belongs to the ML toolbox. For the second task, predicting proficiency levels in mathematics and science (below Level 2, Levels 2–4, or Level 5 and above), another implementation of the RF algorithm was used. For the third task, exploring associations between literacy scores and attitudes towards ICT with the focus on the hierarchical structure of the data, I preferred a traditional statistical method, HLM. In the first section of this chapter, these analytical decisions are justified. Then, results of the study are presented, and methodological recommendations are discussed.

5.1 Analytical Strategy

In this section, I describe my analytical strategy. In terms of method versatility, I applied the toolbox choice approach, as depicted in Figure 5.1.

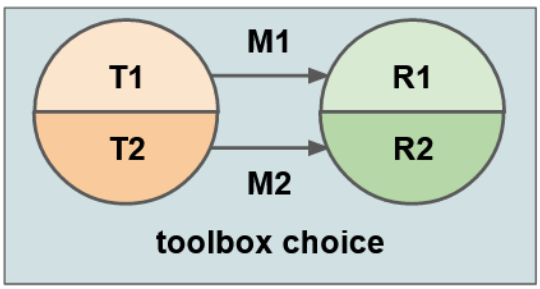

Figure 5.1: **Method Versatility: Toolbox Choice.**
T stands for task, M for method, R for result.

The analytical choices made in this study are related to three consecutive tasks: (a) missing data imputation; (b) the classification task with proficiency levels (below Level 2, Levels 2–4, or Level 5 and above) as the categorical outcome variable, and (c) the regression task with the literacy score as the continuous outcome variable.

Missing data were explored with visualizations from the package VIM [133]. The scatterplot matrices and the aggregation plots were explored. Scatterplot matrices are generalisations of scatterplots to the multivariate case which highlight the missing data in the selected variables. Patterns of missingness can be thus observed. For more details on aggregation plots, see Chapter 6. For missing data imputation, I used the RF algorithm as an effective and unbiased method (see [91], [163], [256]). For its implementation in R, I selected the package missForest [237], as it was shown to be more robust than other versions of RF imputation [245]. Histograms of variables for the complete cases and the dataset after imputation were compared to visually assess the results of imputation. The imputed data were used for further analysis.

For predicting students' proficiency levels in mathematics and science (below Level 2, Levels 2–4, or Level 5 and above), RF classification models were built, and attitudes towards ICT, ESCS and gender were used as predictors. RF was preferred to ordinal regression because RF does not require proportional odds assumption [172] and handles nonlinearity and interactions [69]. I used the package randomForest [145], which is a simple and commonly used implementation of the algorithm based on Breiman's [28] original code. As the sample was imbalanced (there were substantially more students on Levels 2–4), the training set was oversampled with the package UBL [27]. To avoid excessively optimistic performance assessment, I evaluated the models on the test set (20% of the 2015 data) which was not oversampled; and then, on the 2018 data. The AUC was used as a measure of model performance; the multiclass AUC and separate class comparisons were estimated with the package pROC [207]. The mean decrease in accuracy (permutation importance) was chosen as a variable importance measure. Partial dependence plots for each predictor were built with the package pdp [94]. Three-dimensional partial dependence plots for pairs of predictors were built to illustrate their relationships with each other and with the predicted outcome.

In order to obtain a more detailed picture of relationships between attitudes towards ICT and mathematical and scientific literacy with the focus on the nested structure of the data (different schools), a multilevel regression model was required. The RF models as implemented in the randomForest [145] were not suitable due to a limit for factor levels, and a different method to build such a model was needed. It was shown that for predictive multilevel models (such as RF), an increase in the group size is more beneficial than an increase in the number of groups, while for estimative models (such as HLM) the opposite is the case [2]. The group size in this case is the number of student participants in each school, and the number of groups is the number of schools. In both datasets, the group size varied from one to 30 students, while the number of

groups was rather large (254 for the 2015 dataset and 208 for the 2018 dataset), which made HLM the instrument of choice.

HLM was conducted with the package lme4 [17] in accordance with methodological recommendations summarized by Dedrick et al. [53] and Harrison et al. [103]. Restricted maximum likelihood was used as an estimation method [53]. Independent variables were grand-mean centred and standardized by two standard deviations for comparability of continuous and binary variables [83]. The need for multilevel modeling was assessed by exploring variance decomposition in null (unconditional) models with random intercepts [160]. As full models with random slopes and random intercepts, which were applied in accordance with [103], failed to converge, full models with random intercepts were built. To estimate sampling variance, Fay's method with 80 replicates was used, and statistical estimates were averaged for 10 plausible values [178]. As only six predictors were involved in the analysis (gender, ESCS and four ICT attitudes), they were all included in the model simultaneously (see [107] for variable selection methods). Significance of estimates was assessed with the type III Wald test [152], and a significance level of .001 was set as appropriate for these procedures and for the sample size [139]. Nakagawa's marginal and conditional R^2 [166], [167] were calculated with the package performance [151]. Assumptions for residuals were checked with diagnostic plots from the package sjPlot [150]. In check for multicollinearity, the cut-off value of 3 for the variance inflation factor was used as recommended in [273]. ICCs and Nakagawa's marginal and conditional R^2 were reported as effect size measures for random effects [138], [147].

5.2 Data

The data for the study was obtained from the OECD website, where PISA 2015 [176] and PISA 2018 [177] databases are in free access. In this work, German subsets of PISA student questionnaire and the optional ICT familiarity questionnaire for students were used. Cases with 100% missing ICT responses, which were due to the fact that the ICT questionnaire was optional in Germany [203], were removed. In accordance with this criterion, 1093 cases (16.81%) were removed from the 2015 dataset and 944 cases (17.32%) were removed from the 2018 dataset before the analysis. The resulting 2015 sample consisted of 5411 students (50.71% female, n = 2744; 49.29% male, n = 2667) from 254 German schools. The resulting 2018 sample consisted of 4507 students (47.33% female, n = 2133; 52.67% male, n = 2374) from 208 German schools.

The following variables were included in analysis: the student's mathematical and scientific literacy, ESCS, gender, and four attitudes towards ICT. ESCS is a composite score based on the three indicators: parental education, highest parental occupation, and home possessions (which is used as a proxy for family wealth [181]). The measure is constructed via PCA and standardized for a standard deviation of one, with zero representing the overall OECD average [181]. In the datasets, gender was coded as 1 for female and 2 for male; in this analysis, it was recoded as 0 for female and 1 for male. Attitudes towards ICT were measured with the 4-point Likert scale from strongly

disagree to strongly agree, and the following IRT-scaled indices were calculated [182]: (a) perceived ICT competence, based on five items; (b) interest to ICT, six items; (c) ICT in social interaction, five items; and (d) ICT autonomy, five items. The measures were the same in PISA 2018 [183]. Mathematical literacy was defined as 'an individual's capacity to formulate, employ and interpret mathematics in a variety of contexts' [181]. Scientific literacy was defined as 'the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen' [181]. Ten plausible values were included in analysis for each literacy score [178]. In both PISA waves, the cut-off score for performance below Level 2 in mathematics was 420.07 and in science 409.54. The cut-off score for performance at Level 5 and above in mathematics was 606.99 and in science 633.33 (see [184]).

5.3 Results

RF for Missing Data Imputation and Classification Task

In the 2015 dataset, 2.41% of the data was missing, and 3.81% of the data was missing in the 2018 dataset. In Figure 5.2, scatterplot matrices for missing data are presented for the 2015 dataset (the upper matrix) and the 2018 dataset (the lower matrix). In the scatterplot matrices, the red colour depicts missing data, and the blue colour the observed (non-missing) cases. It could be seen that there were no discernible patterns in missingness. This could be also concluded from exploring the aggregation plots for missing values. Thus, the imputation could be conducted, and the imputed data could be used in further analysis without biasing the models.

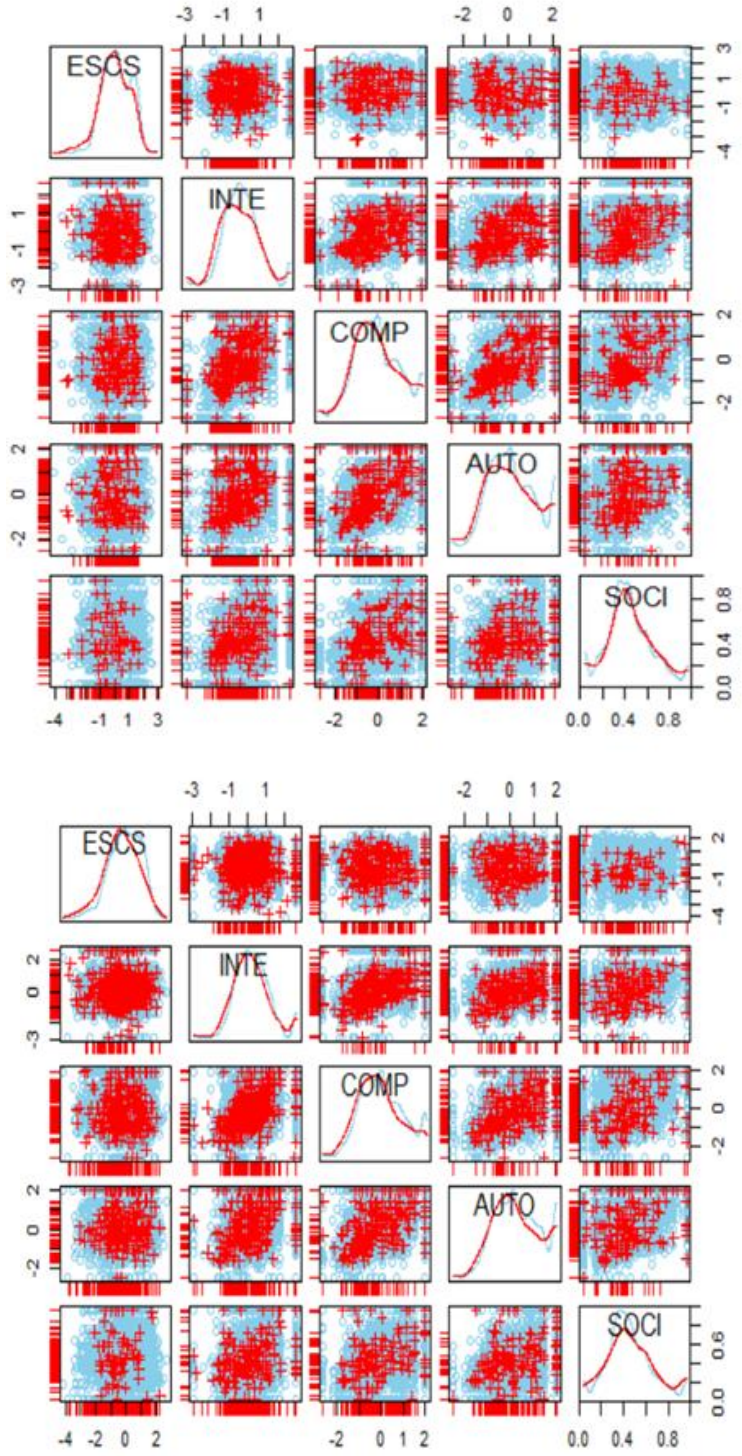

Figure 5.2: **Scatterplot Matrices for Missing Data.**
The upper matrix is for the 2015 data, the lower matrix for the 2018 data. Red is for missing cases, light blue is for observed data. ESCS is economic, social, and cultural status; AUTO is ICT autonomy; INTE is ICT interest; SOCI is ICT in social interaction.

RF imputation was conducted with the default settings (the number of trees = 100, the maximal number of iterations = 10). Histograms of variables from the complete cases dataset and the dataset after imputation (see Figure 5.3) showed that the imputation did not lead to changes in their distributions that could potentially bias the models.

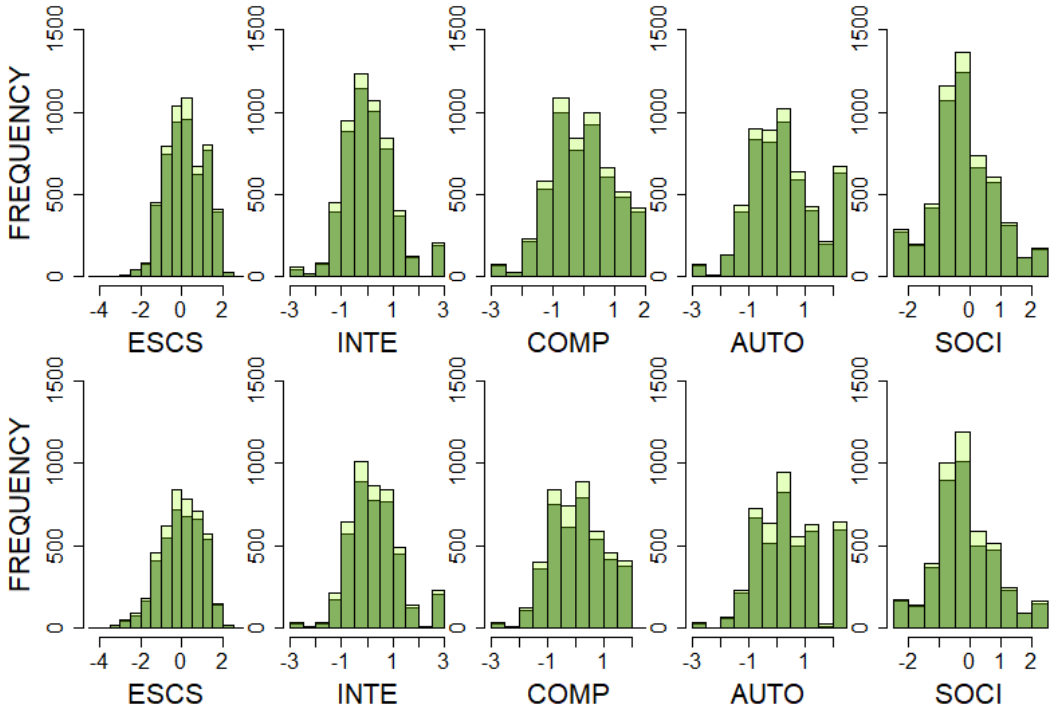

Figure 5.3: **Histograms of Complete Cases and Imputed Data.**
The upper row is for the 2015 data, the lower row for the 2018 data. Dark green is for complete cases, light green is for imputed data. ESCS is economic, social, and cultural status; AUTO is ICT autonomy; INTE is ICT interest; SOCI is ICT in social interaction.

To predict students' proficiency levels (below Level 2, Levels 2–4, or Level 5 and above), two RF models were trained: the mathematics model and the science model. The training set (80% of the 2015 dataset) was used. Variable importance plots for the models are shown in Figure 5.4. Attitudes towards ICT were less important than ESCS but more important than gender in both models, and ICT autonomy was more important than other attitudes.

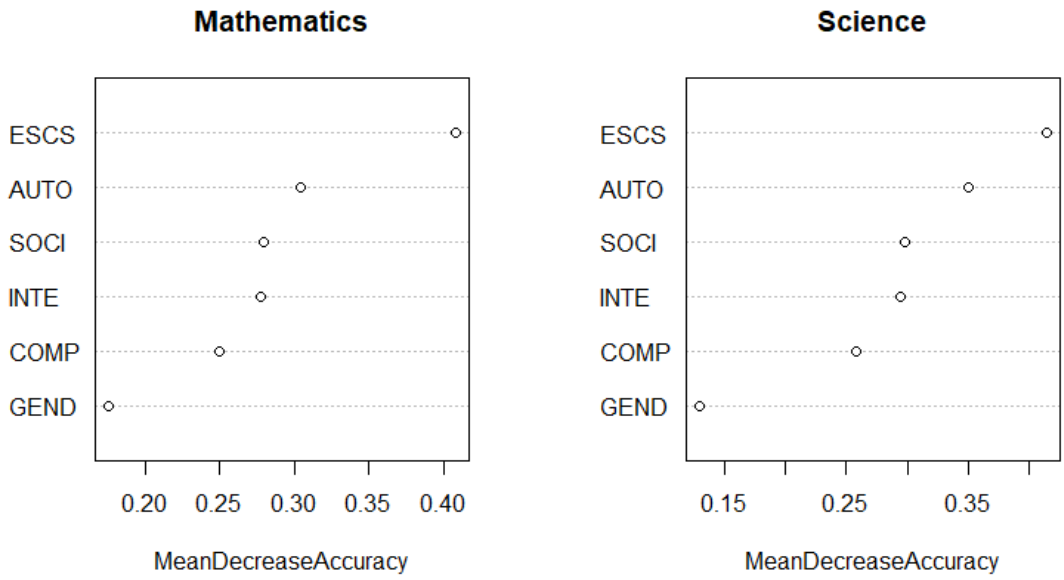

Figure 5.4: **Permutation Variable Importance.**
ESCS is economic, social, and cultural status; GEND is gender; AUTO is ICT autonomy; INTE is ICT interest; SOCI is ICT in social interaction.

Variable importance measures do not indicate whether association of a variable with the predicted outcome is positive or negative; this information can be obtained from partial dependence plots for each variable. In Figure 5.5, the partial dependence plots for the mathematics model and the science model are shown. They reveal rather similar patterns. The plots for class 1 (MATH-1 and SCI-1) indicate the probability for a student to perform below Level 2; the plots for class 3 (MATH-3 and SCI-3) indicate the probability to perform on Level 5 and above.

The probability is predicted by ICT interest, ICT competence, ICT autonomy, ICT in social interaction, and ESCS. In the plots, higher levels of ICT autonomy predicted a higher probability for a student to perform on Level 5 and above in mathematics and science, while lower levels of autonomy increased a probability to perform below Level 2. For ICT in social interaction, the partial dependence plots gave the opposite picture: higher levels of this attitude predicted a higher probability to perform below Level 2. The plots show nonlinearity of relationships between predictors and the predicted outcome, which is most clearly seen for ICT interest.

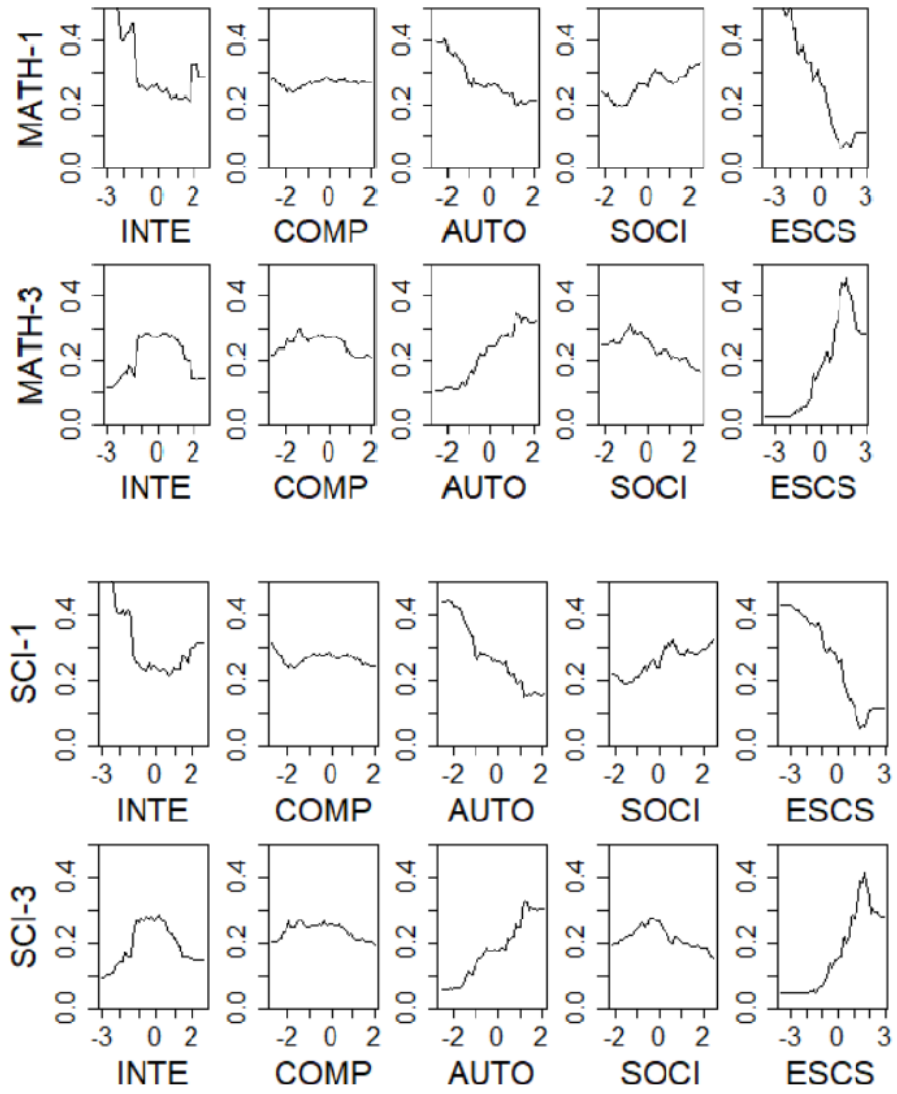

Figure 5.5: **Partial Dependence Plots for Predictors.**
ESCS is economic, social, and cultural status; AUTO is ICT autonomy; INTE is ICT interest; SOCI is ICT in social interaction; MATH-1 is mathematical proficiency below Level 2; MATH-3 is Level 5 and above; SCI-1 is scientific proficiency below Level 2; SCI-3 is Level 5 and above.

Partial dependence plots for pairs of variables indicated nonlinear relationships between them in terms of probability of the predicted outcome for each class (below Level 2, Levels 2–4, or Level 5 and above). In Figure 5.6, partial dependence plots for ESCS and ICT autonomy, the two most important variables, are shown for the science model. It can be seen that the highest probability for a student to perform on Level 5 and above in science was predicted by high levels in both ESCS and ICT autonomy.

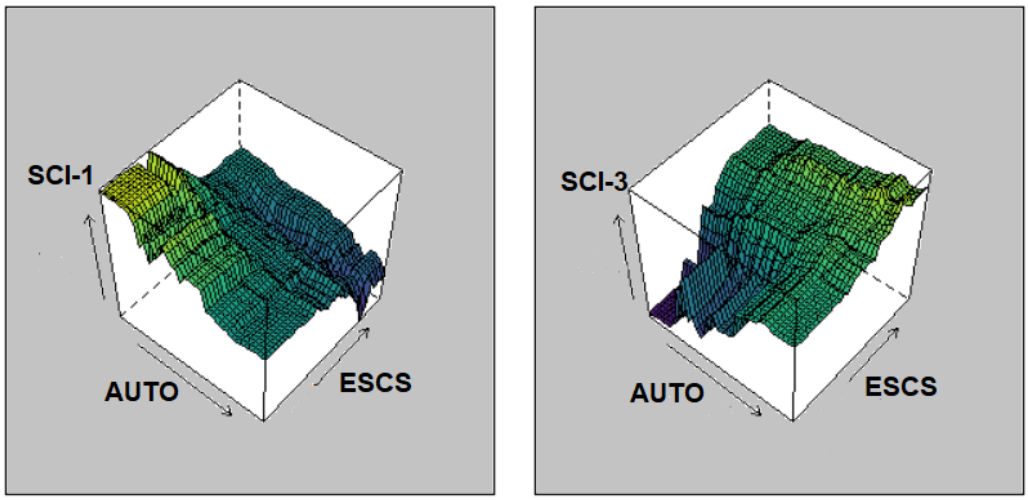

Figure 5.6: **Partial Dependence Plots for the Pair of Predictors.**
ESCS is economic, social, and cultural status; AUTO is ICT autonomy; SCI-1 is scientific proficiency below Level 2; SCI-3 is Level 5 and above.

The models were evaluated on the test set from the 2015 data (20% of the sample). The multiclass AUC was 67.44% for the mathematics model and 71.66% for the science model. When evaluated on the 2018 data (the whole dataset), the multiclass AUC was 69.19% for the mathematics model and 68.51% for the science model. Although the model performance was suboptimal (see limitations in section 8.2), it is noteworthy that it did not change considerably for the 2018 data. The ROC curves for the class comparisons are presented in Figure 5.7.

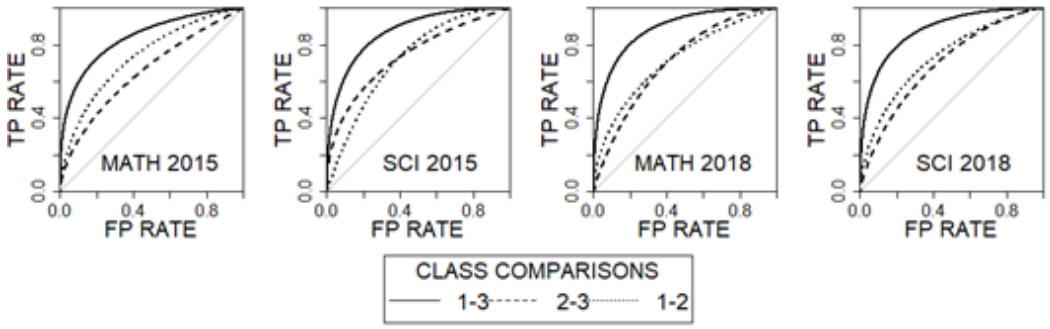

Figure 5.7: **Receiver Operating Characteristic Curve for the Models.**
TP is true positive; FP is false positive; MATH 2015 is the mathematics model fit on the 2015 test set; MATH 2018 is the mathematics model fit on the 2018 dataset; SCI 2015 is the science model fit on the 2015 test set; SCI 2018 is the science model fit on the 2018 dataset; class 1 is proficiency below Level 2; class 2 is Levels 2–4; class 3 is Level 5 and above.

Multilevel Modeling

Unconditional models for mathematical literacy and scientific literacy were built. For mathematical and scientific literacy, multilevel modeling was required, as can be seen from ICC values reported in Table 5.1.

Table 5.1: **Unconditional Models for Mathematics and Science.**

	Mathematical Literacy		**Scientific Literacy**	
	2015	**2018**	**2015**	**2018**
		Fixed effects		
Intercept	506.35***(1.86)	503.76***(1.07)	510.87***(1.24)	506.22***(1.07)
		Random effects		
Intercept variance	3636.65	3822.62	4428.87	4628.37
Residual variance	4541.25	5162.91	5457.53	5956.34
		Effect size		
ICC	.445	.425	.448	.437

In the 2015 dataset, N = 5411; in the 2018 dataset, N = 4507. Standard errors are in parentheses. Fixed effects are parameter estimates. Random effects are variance estimates. ICC = intraclass correlation coefficient.
***p < .001.

For mathematical literacy and scientific literacy, full models with random intercepts were built. Fixed and random effects estimates are reported in Table 5.2. Assumptions for the full models were checked for each of 10 plausible values in mathematical and scientific literacy in the 2015 dataset and the 2018 dataset. Values of the variance inflation factor were below the cut-off value of 3, indicating that there was no multicollinearity in the data. Diagnostic plots for residuals showed that linearity and homoscedasticity assumptions were met, and residuals of the models were normally distributed. As the data was standardized by two standard deviations, it was possible to compare the relative importance of all variables, including binary (gender), based on their regression coefficients.

Table 5.2: **Full Models for Mathematics and Science.**

	Mathematical Literacy		**Scientific Literacy**	
	2015	**2018**	**2015**	**2018**
		Fixed effects		
Intercept	496.16***(1.83)	495.08***(2.08)	505.61***(1.69)	501.20***(1.99)
GEND	22.77*** (2.92)	18.99*** (3.04)	13.78*** (2.26)	12.57*** (3.02)
ESCS	24.59*** (3.00)	33.03*** (3.72)	30.12*** (2.71)	35.97*** (3.62)
COMP	-6.04 (3.63)	-3.15 (3.66)	-6.48 (3.08)	-2.75 (3.84)
INTE	0.15 (3.16)	2.01 (3.11)	-3.63 (2.55)	-0.47 (3.34)
SOCI	-10.76*** (2.88)	-9.21*** (3.18)	-14.33*** (2.88)	-21.08*** (3.25)
AUTO	22.56*** (3.28)	17.45*** (3.41)	38.29*** (2.88)	25.99*** (3.83)
		Random effects		
Intercept var.	2948.67	2897.46	3402.02	3377.05
Residual var.	4218.22	4859.19	5009.35	5614.08
		Effect size		
R^2, marginal	.055	.055	.066	.057
R^2, conditional	.444	.408	.444	.411

In the 2015 dataset, $N = 5411$; in the 2018 dataset, $N = 4507$. Standard errors are in parentheses. Fixed effects are parameter estimates. Gender was recoded as 0 for female and 1 for male. In the 2018 dataset, the relationship between mathematical literacy and ICT in social interaction was nonsignificant at level .001 in three out of 10 models. Random effects are variance estimates. Effect size is Nakagawa's R^2. ESCS is economic, social, and cultural status; GEND is gender; AUTO is ICT autonomy; INTE is ICT interest; SOCI is ICT in social interaction.
***$p < .001$.

ICT in social interaction was significantly negatively associated with mathematical literacy and scientific literacy. ICT autonomy was significantly positively associated with mathematical literacy and scientific literacy, and it was almost as influential as gender and ESCS for mathematical literacy both in the 2015 dataset and in the 2018 dataset. Its association with scientific literacy ($\beta = 38.29$) was the strongest among all variables in the 2015 dataset and the second strongest after ESCS in the 2018 dataset

(β = 25.99). In Figure 5.8, regression coefficients of all variables are shown with confidence intervals in different colours for 10 plausible values.

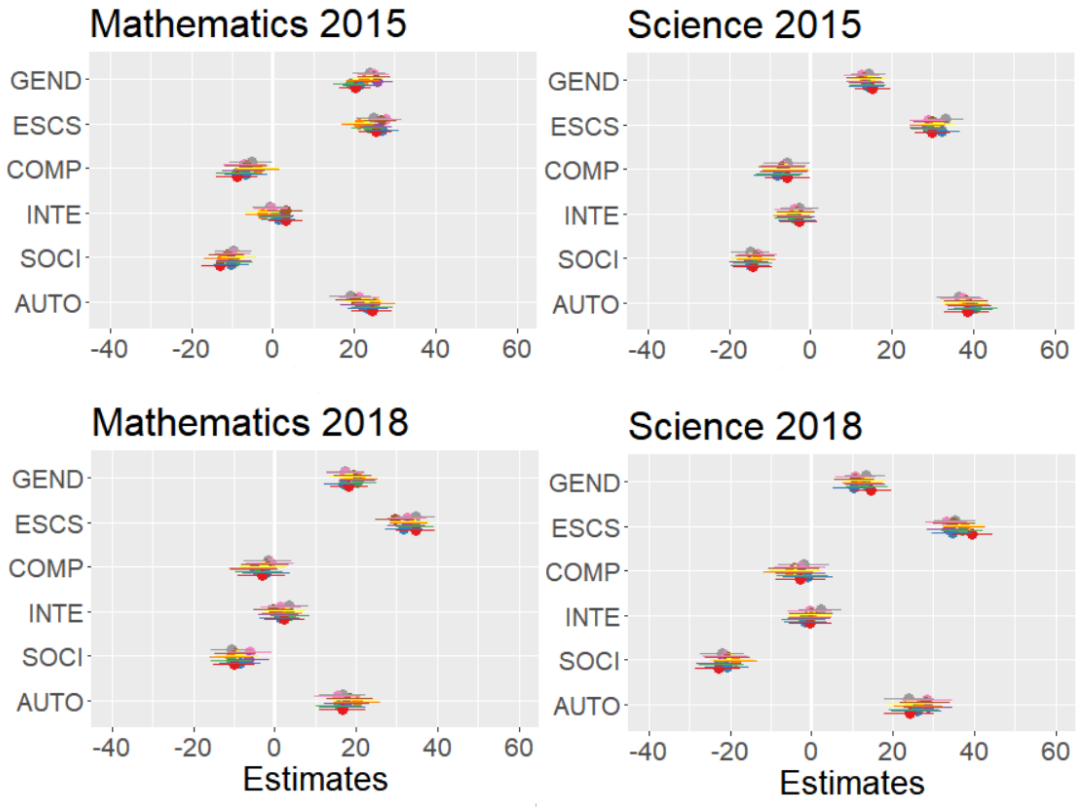

Figure 5.8: **HLM Estimates for Plausible Values.**
ESCS is economic, social, and cultural status; GEND is gender; AUTO is ICT autonomy; INTE is ICT interest; SOCI is ICT in social interaction.

5.4 Summary

In this chapter, I presented a flexible way to combine ML and statistical methods for data analysis of a large-scale educational survey. For each task in the analytical process, I selected the most suitable tool from either statistical or ML repertoire and justified these analytical decisions. In terms of method versatility, it is the toolbox choice of different methods.

From the perspective of educational science, the results of this study highlighted the role of ICT autonomy: for German students, this attitude was an important variable in classification models, and it was significantly positively associated with literacy scores in HLM in PISA 2015 and PISA 2018. These findings, as well as the results on ICT in social interaction (significantly negative associations with mathematical and scientific literacy) and ICT competence (no significant associations), were consistent with previous works [119], [162]. Negative associations of German students' mathematical and scientific literacy with ICT interest in PISA 2015 were reported as significant in

[162] but were non-significant in the current study. This discordance can be explained by the difference in analytical choices; it is to a degree inevitable, as shown in [230], and can be explored in further research. In the publication on the topic, my co-author and I emphasized that the role of ICT autonomy is probably not fully recognized by contemporary educational systems, with their emphasis on collaborative learning and group activities [57]. In line with this tendency, the OECD focuses increasingly on collaborative problem solving in PISA: the change in the optional ICT questionnaire for the next PISA wave, from which any items measuring ICT autonomy were excluded [148], [186], is indicative of this trend. The findings of the current work could be an argument supporting the importance of research on students' ICT autonomy.

From the methodological perspective, I illustrated how the toolbox choice of the most suitable methods from ML and statistical repertoire can benefit data analysis of a large-scale educational survey. The analytical choices reported in this chapter started with exploring and imputing missing data, which is a persistent problem in PISA [114] and in large-scale educational surveys in general [91]. For this task, I chose the RF algorithm implemented in the package missForest based on previous studies showing its effectiveness and unbiasedness (see [163] and [256]) and checked the quality of imputation with histograms of the imputed data in comparison to the complete cases. With a different RF implementation from the package randomForest, I built classification models for scientific and mathematical proficiency levels of German students in PISA 2015 and PISA 2018 in accordance with the three-part division of proficiency levels (below Level 2, Levels 2–4, or Level 5 and above) accepted in PISA (see [180], [184], and [193]). A differently formulated classification task, such as predicting the whole range of proficiency levels, or low vs regular academic performance (a binary task), would have been also possible with these data. While choosing an instrument for the classification task, I preferred RF to ordinal regression [172] and to more sophisticated ML models (see section 8.2). Performance of the models built for the 2015 data did not decrease on the 2018 data, indicating that the same patterns persisted in both years. With the help of model-agnostic methods, I assessed the relative importance of the variables (attitudes towards ICT, ESCS, and gender) for proficiency levels and visually presented nonlinear relationships between the predictors and the predicted outcome. To explore relationships between attitudes towards ICT and mathematical and scientific literacy in more detail, a multilevel model that takes into consideration the hierarchical structure of the data was needed. For this specific analysis, because of the dataset characteristics [2], HLM from a statistical toolbox was preferable to predictive ML models. With HLM, I obtained information on fixed and random effects, such as significance of relationships, their direction, and effect sizes. The 2018 data revealed the same patterns in associations between ICT attitudes and literacy scores as the 2015 data, and fixed effects estimates in HLM for these two years did not substantially differ.

Combining ML and statistical approaches is beneficial for research on large-scale educational surveys, as the former is a valuable tool for finding generalizable patterns, while the latter is useful for testing hypotheses and making statistical inferences. This

chapter showed that a flexible choice of analytic instruments from both toolboxes depending on the study aims and the dataset characteristics is an effective way of analysing the data.

Chapter 6

Selecting the Number of Clusters in Latent Class Cluster Analysis

This chapter continues the topic of method versatility at the stage of research *per se* in relation to data analysis of large-scale educational surveys, which was discussed in the previous chapter. In Chapter 5, I showed how method versatility can be introduced into supervised learning methods applied in the area by combining ML and statistical toolboxes to analyse PISA data. In this chapter, I explore versatility in unsupervised learning methods of educational data analysis to address the same RQ2, which I repeat here:

> **RQ2.** How to facilitate method versatility in educational research on human attitudes towards technology?

This chapter describes method versatility in unsupervised learning (cluster analysis) applied to large-scale educational surveys. I show that the simultaneous use of different criteria for cluster selection in LCCA leads to detecting separable generalizable clusters in the data. As I have already mentioned, novel clustering algorithms are often developed, from the perspective of method versatility, to extend the range of existing tools. Here, the aim is different: I do not present a novel clustering method but suggest a strategy for a more effective use of existing approaches. LCCA deals with unequal covariance matrices, unequal number of observations in the cluster, and poorly separated clusters, which are typical for real-world datasets, and thus this model-based clustering method is useful for large-scale educational surveys.

As this thesis focuses on the topic of human attitudes towards technology, I illustrate the suggested strategy on the dataset of German teachers' attitudes towards ICT from ICILS 2018. Studies using ICILS data mostly focus on students [32]; LCCA on teachers' attitudes to ICT in ICILS 2013 was conducted in [63]. In this study, the number of clusters was selected based solely on model fit. It is a common procedure in the area, and the name "model-based clustering" already implies that model fit indices

play a pivotal role in the selection. However, it was shown that criteria such as the ASW used in distance-based clustering can aid the selection, and considering the stability of cluster partitions leads to selecting generalizable clusters. Therefore, in this chapter I suggest a strategy integrating model fit, cluster separation, and the stability of partitions criteria. To illustrate the strategy, I apply it to the simulated data with the known cluster structure and to the real world dataset from ICILS 2018 [78]. With the simulated data example, I show how model fit and cluster separation could be considered in terms of the trade-off between them. With the ICILS data, I provide end-to-end LCCA starting with preprocessing of the data and including all steps of the selection procedure.

6.1 Analytical Strategy

The analytical strategy contains explanation of choices regarding data preprocessing, conducting LCCA, selecting the number of clusters, and visualizing the results. The cluster selection involves criteria of model fit (the BIC and the ICL), cluster separation (the ASW), and the stability of partitions, with the parsimony and interpretability considerations taken into account. In terms of method versatility, it is the simultaneous use of different methods (Figure 6.1). The data analysis was conducted with R, version 4.0.2 [196]. The R script is available on GitHub https://github.com/OlgaLezhnina/LCCA.

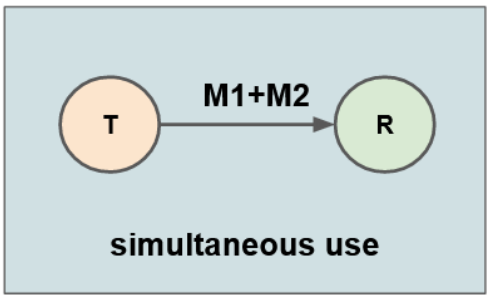

Figure 6.1: **Method Versatility: Simultaneous Use.**
T stands for task, M for method, R for result.

Prior to the analysis, a few preprocessing procedures should be conducted. Firstly, a hierarchical structure of the data needs to be explored to decide whether LCCA is sufficient, or multilevel LCA is needed based on the ICC values [165]. Then, missing data should be explored, and an imputation procedure chosen. The aggregation plot for visualizing missingness and detecting possible patterns in missing data could be built with the package VIM [133]. For imputation, I used the RF algorithm as an effective and unbiased imputation method [91]. Variable selection is an important step of the data preprocessing [75], but it was not applied in this work, as the explored models included all variables of interest. Normalization of variables is not required for LCCA, so this step, common for other techniques, could be omitted. Dichotomization of response options, although not infrequent in LCCA research [63], might be considered

objectionable [154]. Therefore, the decision on dichotomization was made based on exploring the frequencies of endorsements of different answer options.

The strategy for selecting the number of clusters in LCCA includes criteria based on model fit, cluster separation, and the stability of partitions. In order to provide the researcher with detailed information on model fit and cluster separation, I wrote an easy-to-use custom function (LCCAselection) based on the VarSelClust function from the VarSelLCM package [158]. The function returns a data frame with information criteria and silhouette indices for one- to ten-cluster solutions. As visualizations tools were shown to be important for deciding on the number of clusters (see [74], [95]), the graphical output is included in the custom function to aid the cluster selection. The function produces a plot that integrates (i) the BIC plot for all cluster solutions, so that the elbow heuristic can be applied, (ii) the ASW plot for all cluster solutions, and (iii) vertical lines indicating the minimal BIC and the minimal ICL. Thus, the researcher can make informed decisions regarding model fit and cluster separation. When two or three best solutions are selected, their stability can be checked by the other custom function (valfunc). This second function accepts the data, the number of clusters, and the number of bootstrap samples as arguments to return the Jaccard coefficient and the ARI for bootstrap validation of the cluster solution. These coefficients were chosen as they are two most widely used and easily interpretable metrics [109]. Other considerations, such as parsimony, the size of population shares, and the interpretability of clusters need to be taken into account for the final choice of the number of clusters. The most parsimonious cluster solution is preferable in case it satisfies other requirements, and clusters with excessively small population shares are considered inadequate regardless of the fit of the model [190]. Clusters should be interpretable from the perspective of domain knowledge of the researcher.

The selected clusters can be explored and visualised. For this purpose, I presented the PCA visualisation, the silhouette plot for clusters, the barplot for the discriminative power of the variables, and the item probability plot. The PCA presents the clustered data in the two-dimensional projection and can be thus misleading; the values of silhouette widths in the silhouette plot are more reliable indicators of cluster separation. The discriminative power of the variables is defined as the logarithm of the ratio between the probability that the variable is relevant for the clustering and the probability that the variable is irrelevant for the clustering, given the best partition [158]. The greater value indicates that the variable is more important for the clustering. The item probability plot graphically presents item-response probabilities for the selected clusters, or latent classes.

6.2 Data

6.2.1 Simulated Data

The ordinal clustered data was simulated by generating the metric data and applying the discretization process to each variable with the clusterSim package [255]. The easily reproducible R code for simulating and analysing the data can be found at https://github.com/OlgaLezhnina/LCCA. The datasets contained the known structure of clusters. Three datasets with four clusters (N = 1550) and three datasets with six clusters (N = 2250) were generated, each with four response categories and six variables. The six datasets were named A-F. As the influence of the number of variables, the number of categories, sample size and unequal cluster sizes on LCCA performance was explored in large-scale simulation experiments [7], in this work I focused on cluster separation issues relevant to selecting the number of clusters. Therefore, the clusters had unequal covariance matrices and unequal number of objects in them, which is typical for real-world data. The number of "true" (formally assigned) clusters did not coincide with the number of separated clusters in datasets A, B, D, E. The values of silhouette widths for all clusters are presented in Table 6.1.

Table 6.1: **Simulated Clustered Data.**

Dataset	Clusters	Cluster silhouette widths	Cluster samples
A	4/1	.11, –.01, –.03, –.10	600, 200, 500, 250
B	4/3	.33, .26, .26, .09	600, 200, 500, 250
C	4/4	.86, .82, .81, .94	600, 200, 500, 250
D	6/4	.80, –.29, .81, .93, –.24, .34	600, 200, 500, 250, 400, 300
E	6/4	.72, .02, .65, .86, –.32, .76	600, 200, 500, 250, 400, 300
F	6/6	.61, .78, .57, .59, .56, .58	600, 200, 500, 250, 400, 300

The number of clusters is given as total/separated.

6.2.2 The ICILS Dataset

The dataset of teachers' positive and negative views on ICT from ICILS 2018 was retrieved from the IEA website, where it is in free access for downloading [123]. The sampling method for the ICILS teacher survey consisted in randomly selecting 15 teachers from those who teach regular school subjects to the students in the target grade (generally, grade 8) at each sampled school [76]. For this study, I selected the German subset of the data.

In ICILS 2018, teachers' views on using ICT for teaching and learning was measured by asking them whether they agree ("strongly agree," "agree," "disagree," or

"strongly disagree") with a number of statements. Seven of these statements referred to positive results of using ICT in education (positive views), and another six statements to potential impediments of learning (negative views). The scores were on Likert scale from 1 (strongly agree) to 4 (strongly disagree) [77]. For this analysis, the positive views scores were recoded (reversed), so that higher scores represent more positive attitudes to ICT. Prior to the analysis, 57 rows with 100% missing variables were removed (.024 of the dataset), as these participants did not give responses to any of the items. The resulting sample consisted of $N = 2271$ teachers from 182 German schools.

6.3 Results

6.3.1 Simulated Data: Model Fit and Cluster Separation

With the simulated data example, I illustrated the trade-off between model fit and cluster separation, as the cluster structure of the data, in terms of total number of clusters versus the number of separated clusters, was known, and the results could be thus assessed. The LCCCAselection function was evaluated as a possible aid to the researcher who is aiming to select clusters in LCCA that are not only feasible from the model fit perspective but also well-separated.

The LCCAselection function was applied to the six simulated datasets. It appeared, as shown in Figure 6.2 and Figure 6.3, that the minimal BIC (vertical dotted lines in the plots) tended to indicate the "true" number of clusters in the data. The minimal ICL (dot-dashed lines) favoured well-separated clusters in datasets B and D, but not in dataset E. The BIC elbow heuristic, together with the maximal ASW, indicated the number of separable clusters in all datasets: one cluster ("non-clusterable" data) for dataset A, three for dataset B, four for datasets C, D, and E, and six clusters for dataset F. For well-separated clusters in the dataset C, all indices coincided in pointing at the correct cluster solution, and for the dataset A, the criteria were able to detect the problem of non-separated clusters.

The most interesting situations were presented by datasets B, D and E, in which the number of "true" clusters did not coincide with the number of separable clusters. In these cases, the fit indices and the elbow heuristic indicated different cluster solutions.

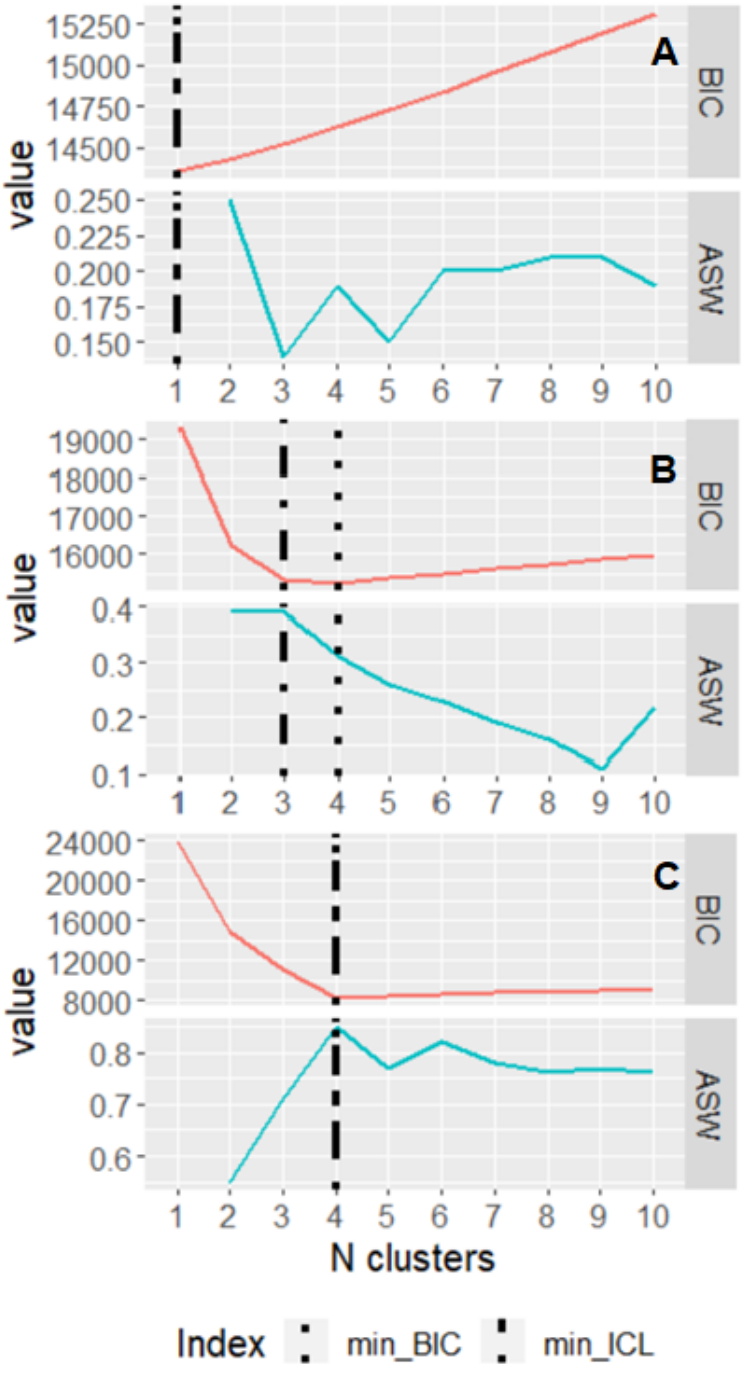

Figure 6.2: **Results of LCCAselection Function on Simulated Datasets A, B, C.**

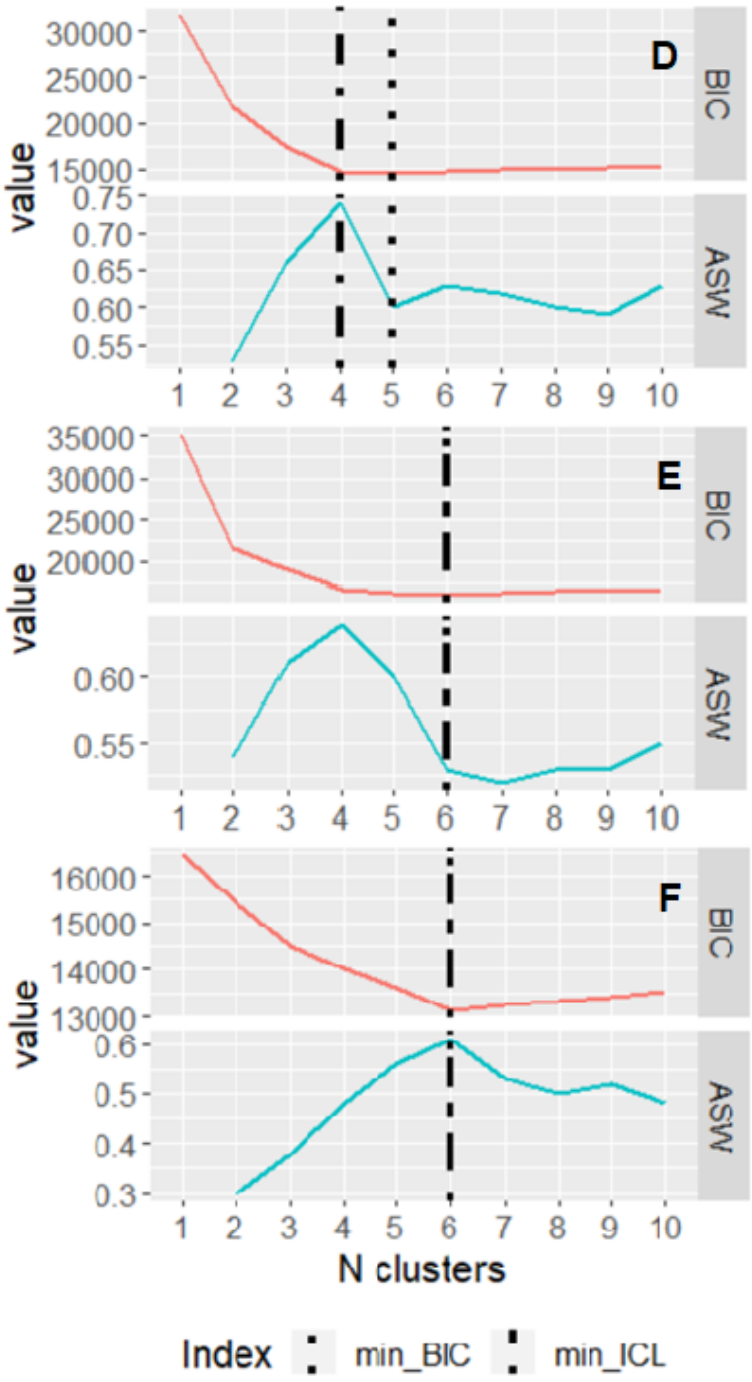

Figure 6.3: **Results of LCCAselection Function on Simulated Datasets D, E, F.**

This example shows how the trade-off between model fit and cluster separation works in LCCA. For instance, for dataset E, if the minimal BIC and the ICL solution is chosen, the ARI = .91 and the ASW = .53. If the BIC elbow solution is preferred, it will

result in better separated clusters with the ASW = .64, but there will be a decrease in the ARI = .73 (see Table 6.2).

Table 6.2**: Cluster Selection Criteria (Simulated Data).**

Dataset	Min BIC		Min ICL		BIC elbow plus ASW	
	ARI	ASW	ARI	ASW	ARI	ASW
A	—	—	—	—	—	—
B	.80	.31	.67	.39	.67	.39
C	1	.85	1	.85	1	.85
D	.65	.60	.51	.74	.51	.74
E	.91	.53	.91	.53	.73	.64
F	1	.61	1	.61	1	.61

Min BIC is the minimal Bayesian Information Criterion; BIC elbow is the elbow heuristics for the BIC plot; min ICL is the minimal Integrated Completed Likelihood criterion; ASW is Average Silhouette Width; ARI is Adjusted Rand Index.

When the researcher aims for compact and well-separated clusters, the BIC elbow heuristics with the maximal ASW might be preferable to the minimal BIC value. Thus, the LCCAselection function can provide an aid in finding well-separated clusters in the data and making informed decisions about balancing model fit and cluster separation criteria. In this brief vignette, I did not deal with the stability of partitions; end-to-end LCCA on the real-world data is presented in the next section.

6.3.2 ICILS Data: End-to-End LCCA

The analysis was conducted on the ICILS teachers' positive and negative views datasets (items A-M). The hierarchical structure of the data was explored. Multilevel ICCs for variables were from .002 to .039, and thus, non-multilevel methods could be used. Missing data (.01 of the dataset) was explored and visualized with the aggregation plot (Figure 6.4).

In the aggregation plot, light blue colour indicates observed data, and dark red colour is for missing data. The barplot in the upper part of the panel shows the proportion of missingness in each variable; it can be seen that it ranges from .05% to 3%, which is very small and implies that the imputation is viable. The grid in the lower part of the panel shows combinations of missing and observed data, and the horizontal bars to the right of the grid show the frequencies of these combinations. Patterns of missingness that would imply that the data was missing not at random were not detected. The scatterplot matrices for missing data (see the description of this visualization in Chapter 5) confirmed this conclusion. Imputation was conducted with

the RF algorithm (the number of trees = 100, the maximal number of iterations = 10), and the resulting dataset was used for the further analysis.

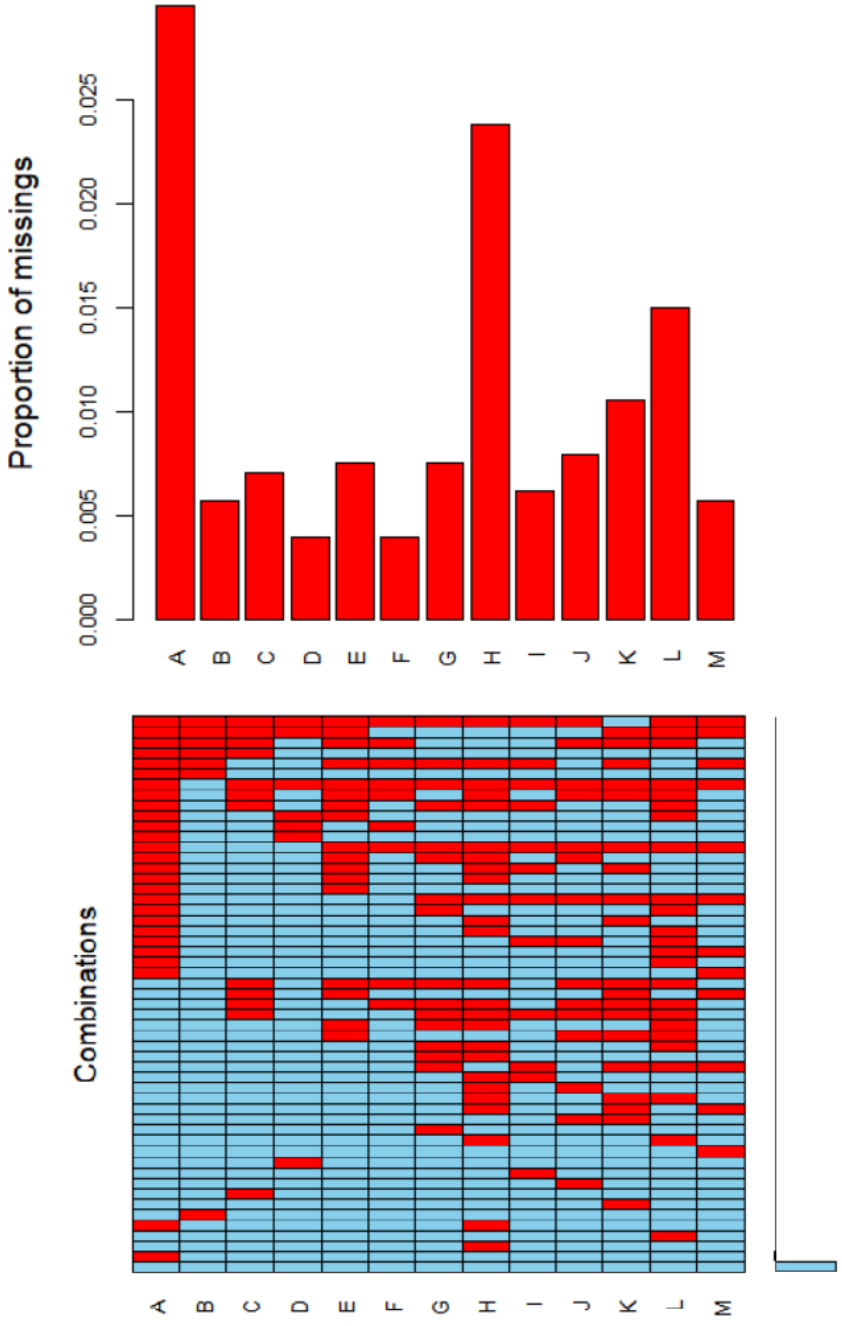

Figure 6.4: **Aggregation Plot for Missing Data.**

Frequencies of endorsements of different answer options for each item were explored (Figure 6.5). The extreme options ("strongly agree" and "strongly disagree") were not underrepresented, and merging them with "agree" and "disagree" would lead to a substantial loss of information. Thus, it was preferable not to dichotomize the data.

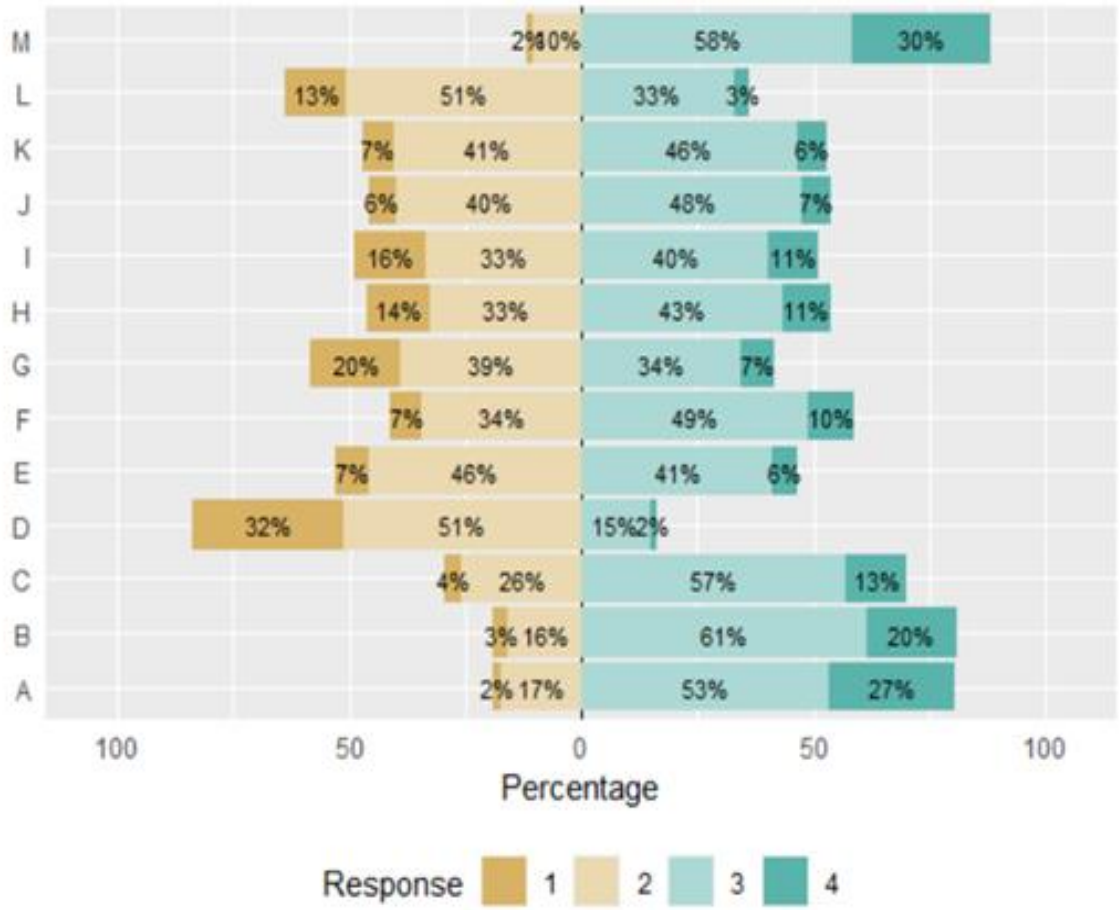

Figure 6.5: **Frequencies of Endorsement for the Items.**

The custom function LCCAselection was applied to the positive views dataset (items B, C, E, J, K, L, and M) and the negative views dataset (items A, D, F, G, H, and I) to select the number of clusters. For the positive views dataset, the BIC elbow heuristic and the maximal ASW pointed at the four-cluster solution, while the minimal BIC indicated the six-cluster solution. The minimal ICL, though, pointed at the seven-cluster solution. For the negative views dataset, all criteria indicated the four-cluster solution.

In Figure 6.6, the graphical output of the LCCAselection function for the negative views dataset and the positive views dataset is presented. The positive views dataset is labelled ICIL-P, and the negative views dataset ICILS-N. Further analysis is presented only for the positive views dataset, and analysis of the negative views dataset was left to the interested reader.

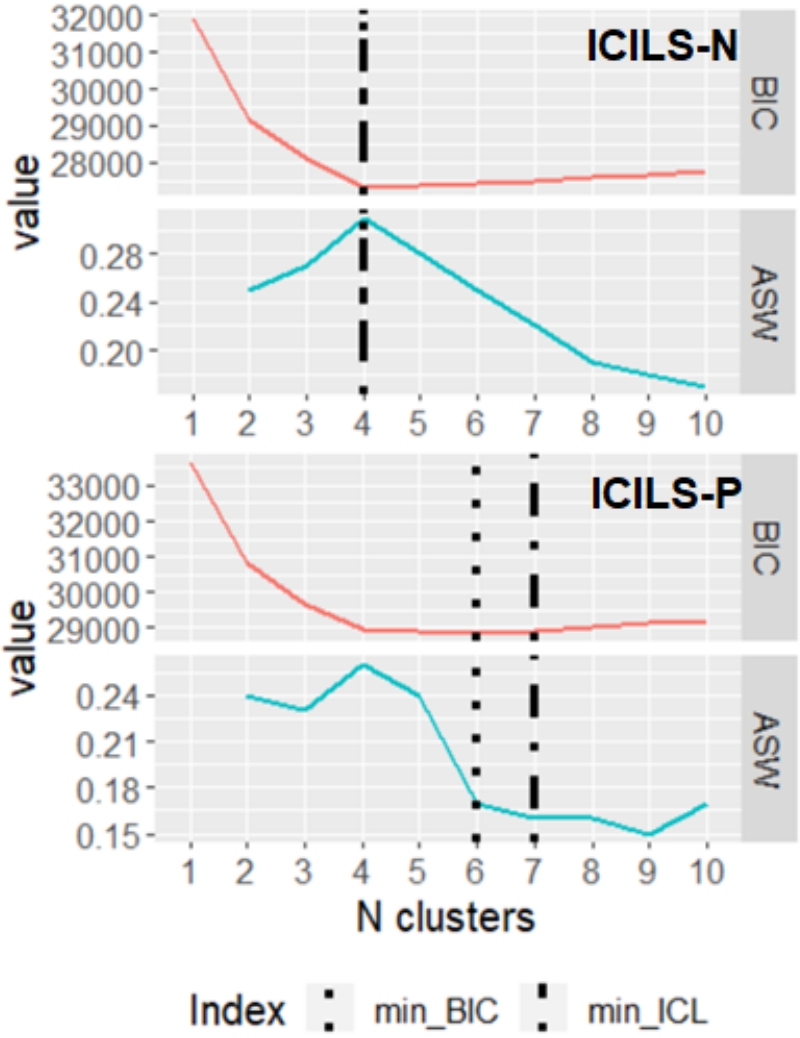

Figure 6.6: **Results of LCCAselection Function on ICILS Data.**

The values of the cluster selection criteria for the ICILS-P data are reported in Table 6.3. It can be seen that for the seven-cluster solution, which was indicated by the minimal ICL, the ASW value (.16) was lower than for the four-cluster solution (.26) indicated by the maximal ASW plus the elbow heuristic, or the six-cluster solution (.17) indicated by the minimal BIC. It meant that the seven-cluster solution had to be rejected, which was also in accordance with the parsimony considerations.

Table 6.3: **Cluster Selection Criteria (ICILS-P Data).**

N clusters	BIC	ICL	ASW
1	33639.57	–16816.52	—
2	30843.22	–15598.33	.24
3	29648.88	–14998.36	.23
4	28985.07	–14684.43	.26
5	28902.88	–14655.55	.24
6	28880.90	–14693.07	.17
7	28925.91	–14651.72	.16
8	29016.39	–14767.68	.16
9	29115.62	–14736.12	.15
10	29217.73	–14804.14	.17

BIC is Bayesian Information Criterion; ICL is Integrated Completed Likelihood criterion; ASW is Average Silhouette Width.

The four- and the six-cluster solutions were compared as shown in Table 6.4. Bootstrap validation with 100 bootstrap samples was used. For the four-cluster solution, the ARI was .88 and the Jaccard index .85, while for the six-cluster solution, the ARI was .76 and the Jaccard index .70. Thus, the four-cluster solution was more stable. In addition, the six-cluster solution had a very low population share in one of the clusters (.03 of the sample). Therefore, the stable and parsimonious four-cluster solution was selected as the final cluster model.

Table 6.4: **Four- and Six-Cluster Partitions Compared.**

N clusters	ARI	Jaccard	Cluster shares (min-max)
4	.88	.85	.10 -.43
6	.76	.70	.03 - .41

ARI is Adjusted Rand Index.

In the upper part of Figure 6.7, the results of PCA for the four-cluster solution are visualized. Such visualisations can be misleading, as multidimensional data is presented in a two-dimensional projection. Therefore, the lower part of the figure shows the silhouette widths for the four selected clusters. The ASW of the four clusters is .26, with the widths of the clusters ranging from .13 to .32. The results show that although the most separable clusters were found in the data, they still cannot be considered perfectly separated, and LCCA as the method of choice was preferable to distance-based methods.

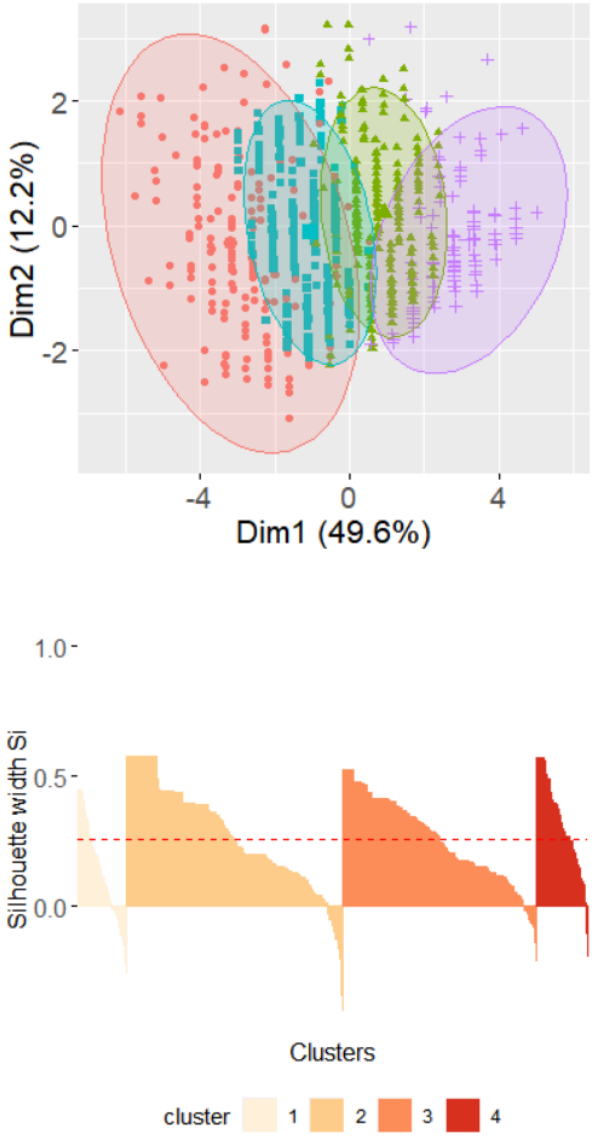

Figure 6.7: **Cluster Visualization and Silhouette Plot.**

Discriminative power of variables was calculated and presented in Figure 6.8. The greater values indicate that the variable is more important for the clustering. Thus, it can be seen that item M was the least informative for the model. Item M is formulated as follows: “The use of ICT in teaching and learning enables students to access better sources of information”.

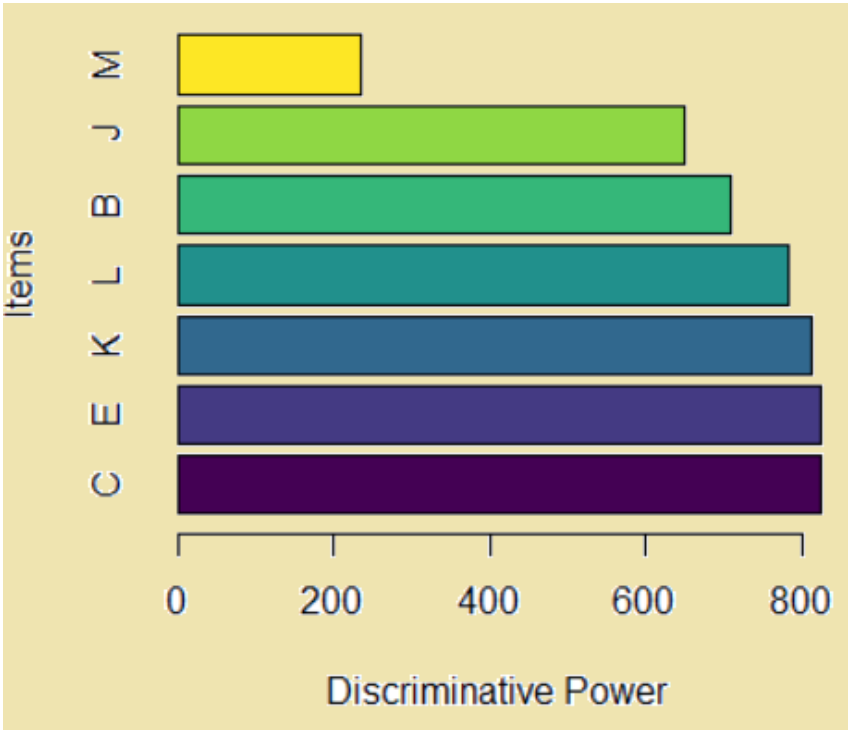

Figure 6.8: **Discriminative Power of Variables.**

The item probability plot for the selected solution was built. The order of the classes was changed to convey the ordinal information.

In the first class (10% of the sample), which could be described as “ICT-enthusiastic”, most participants tended to agree or strongly agree with statements conveying positive attitudes towards ICT in teaching and learning. The second class (43% of the sample), which can be called “ICT-accepting”, also agreed with the statements, although less frequently endorsed the option “strongly agree”. The agreement to some degree persisted in the third class (37% of population), which can be described as “ICT-cautious”, and in which the response “disagree” to most items prevailed. The fourth class (10% of the sample), which might be labelled “ICT-sceptical”, disagreed or strongly disagreed with most items. In particular, the response “strongly disagree” prevailed for item L, which is formulated as follows: “The use of ICT in teaching and learning improves academic performance of students”.

Item M, which had, as reported previously, the least discriminative power in this clustering model, was positively endorsed by most representatives of all four classes,

and thus could not be used for differentiating the classes. The item probability plot, which visualizes these results, is presented in Figure 6.9.

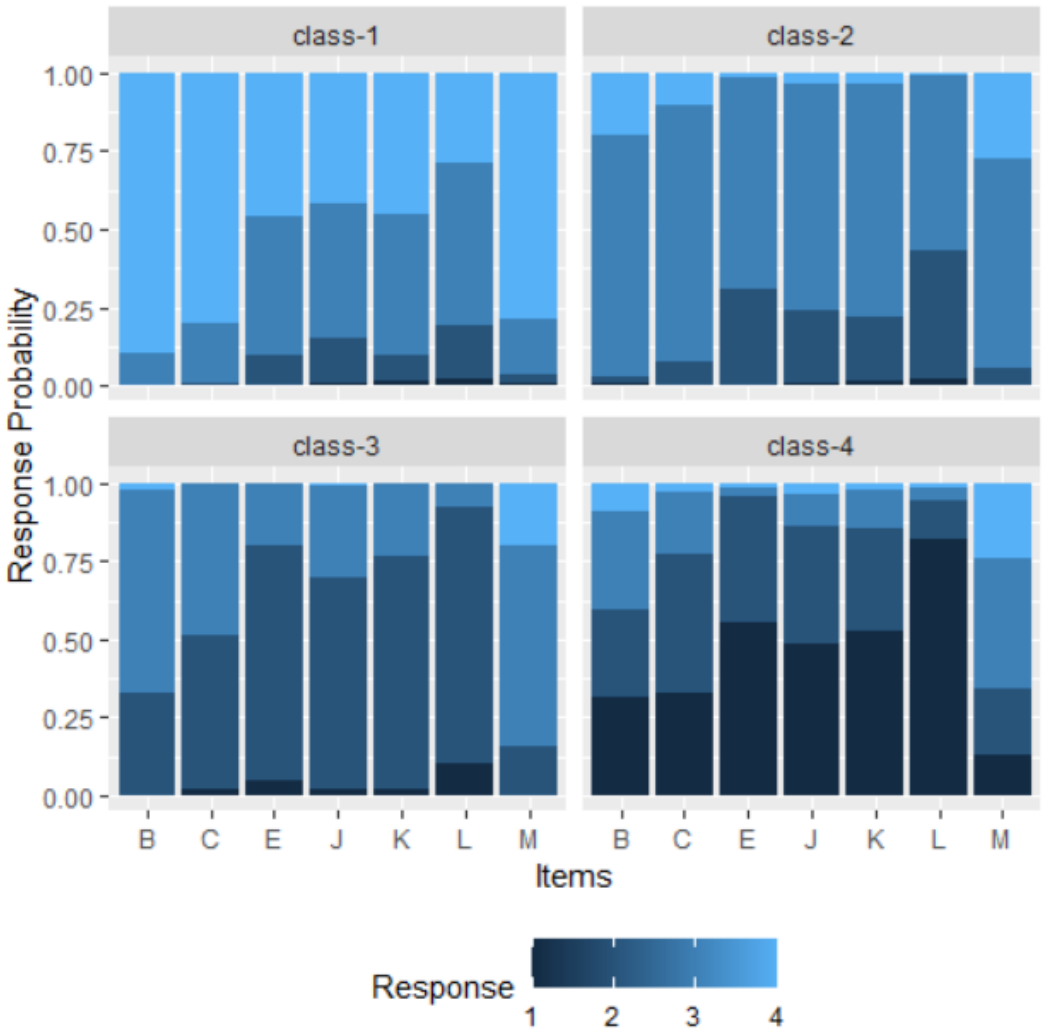

Figure 6.9: **Item Probability Plot.**

6.4 Summary

In this chapter, I employed criteria of model fit, cluster separation, and the stability of partitions to select separable generalizable clusters in the data. In terms of method versatility, it is the simultaneous use of different methods.

From the perspective of educational data analysis, I provided the researchers with end-to-end LCCA including preprocessing of the data, selecting the number of clusters, and interpreting the results. Generalizable separable clusters were selected, interpreted, and visualized in the data of German teachers' positive views on ICT from ICILS 2018. Conclusions about the items and their role in the model (for instance, the least

informative item M) can be useful for the researchers dealing with the ICILS data or conceiving a new scale measuring teachers' attitudes towards ICT.

In methodological terms, I showed that the suggested strategy is more comprehensive than a selecting strategy based on fit indices, be that the most commonly used BIC or the ICL that intends to find separable clusters. In the simulated data example, the combination of the fit indices and the ASW gave the clearest picture of the separable clusters. In case of the ICILS-P data, the strategy led to finding the optimal four-cluster solution, while overreliance on fit indices could have resulted in the choice of the six- or seven-cluster solution that would be less beneficial in regard to cluster separation and the stability of partitions.

The strategy suggested in this chapter widens the scope of tools for conducting LCCA. With a few easily reproducible steps, the researcher can select a cluster solution with optimal model fit, cluster separation, and the stability of partitions. Thus, generalizable interpretable clusters can be more effectively found in the data.

Chapter 7

The ORKG Dashboard: Development and Evaluation

In the previous chapters, I addressed RQ1 and RQ2 and dealt with method versatility at the measurement stage and the research *per se* stage of the research cycle. I explored various ways of facilitating method versatility in data analysis, firstly, in relation to validation of a psychometric instrument, and secondly, when applying supervised and unsupervised learning methods to educational data. In Chapters 4, 5, and 6, method versatility was introduced as the consecutive use, the toolbox choice, and the simultaneous use of different methods, respectively. This chapter is related to the communication stage of the research cycle and addresses the third research question:

RQ3. How to facilitate method versatility in communication of research results related to human attitudes towards technology?

In this chapter, I describe development and evaluation of the ORKG-powered dashboard. As I explained in section 3.3, an effective and fairly intuitive interface was already implemented in the frame of the ORKG research service infrastructure initiative. The ORKG resource comparison enables the user to complete the tasks formulated in [26], that is, to get research field overview; to find related work; to assess relevance; to extract relevant information; to get recommended articles; to obtain deep understanding of the topic; and to reproduce results. My idea was, in terms of method versatility, to extend the range of means for scholarly communication. An alternative interface, which can complement the existing one, might be useful for widening the audience that accepts SKGs as the novel method of scholarly communication. My intention was not to outmatch the ORKG resource comparison in terms of usefulness and performance but rather to create a visualisation that can reach a different section of the audience. The new interface might become a useful addition to the existing one in certain tasks. As explained in section 2.4, a dashboard, which maps the textual modality to the visual modality [157], can be effective for widening the range of communication tools. In comparison to other visualisations, such as mapping graphene research [248]

or science citation knowledge extractor [143], the dashboard was supposed to present a more limited range of contributions because it was designed as an experimental interface for the purpose of preliminary assessment. In my work towards this goal, I based the development of the ORKG dashboard on the principles of technology acceptance outlined in section 2.5 and conducted an evaluation survey to assess the perception of the new service by the potential users.

7.1 Dashboard Development

This section presents the development of the ORKG dashboard. In terms of method versatility, it is the range extension as shown in Figure 7.1.

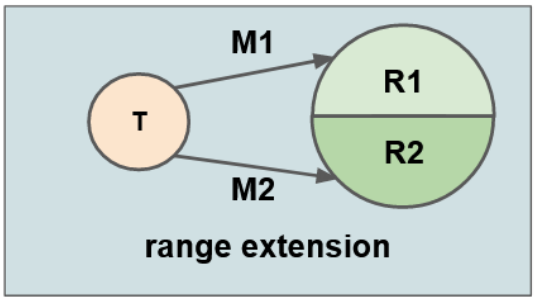

Figure 7.1: **Method Versatility: Range Extension.**
T stands for task, M for method, R for result.

The code for running the dashboard locally is in Python, with the template in HTML, CSS and JS. The code for data analysis of the user evaluation survey is in Python. The participants' responses are publicly available as a CSV file; the answers to open questions were removed due to the data protection considerations. The code and the data can be found at:

https://github.com/OlgaLezhnina/dashboard
https://github.com/OlgaLezhnina/dashboard_survey

The system architecture is shown in Figure 7.2. The backend is a Flask server (Python), with the orkg library used for queries and the pygal library used for visual presentation of results. To generate a webpage, the backend queries the ORKG server to get all information required for the scope of the dashboard; the results are embedded into the generated webpage using Jinja templates. Any other operations are handled by the JavaScript frontend. When the users interact with the dashboard interface, the information stored in the webpage is queried and displayed to them. The presentation is dynamic, and when a new paper on the topic is added to the ORKG, the dashboard contents are automatically updated.

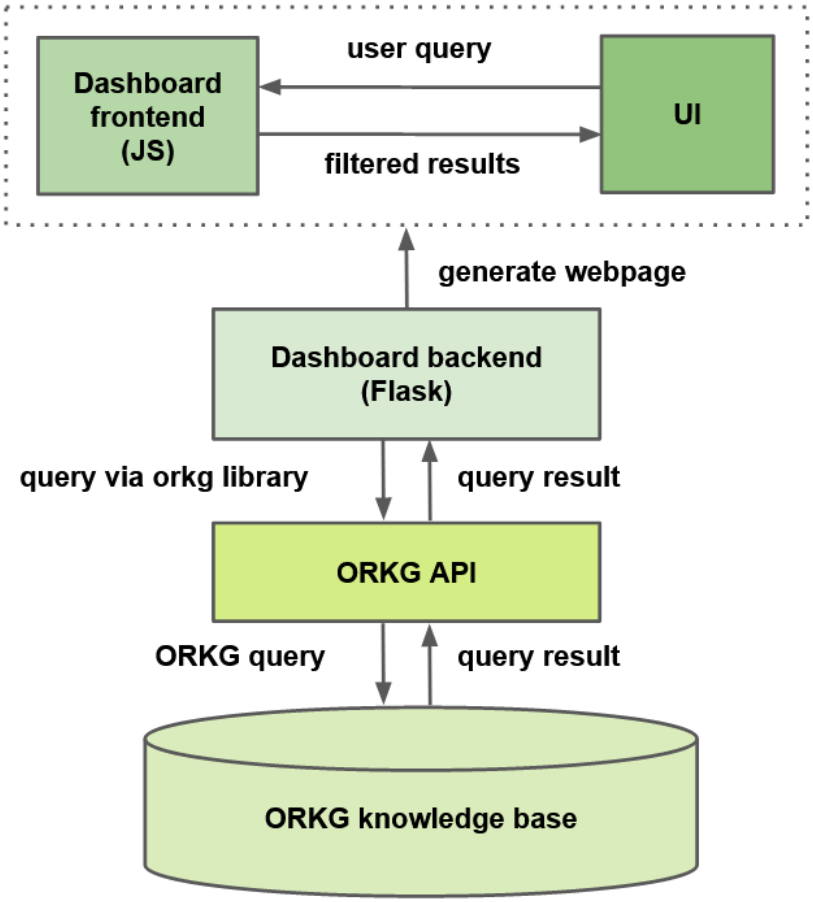

Figure 7.2: **System Architecture of the Dashboard.**
API is application programming interface, JS is JavaScript, ORKG is Open Research Knowledge Graph, UI is user interface.

In terms of the topic of academic papers that had to be added to the ORKG as resource contributions and then visualized with the dashboard, there were no specific requirements. Therefore, I selected the topic of students' attitudes towards ICT in PISA 2015 and 2018, which was explored in depth in Chapter 5.

This choice of the topic was determined by the following considerations:

(i) The topic belongs to the area of human attitudes towards technology and is therefore relevant to the overarching concept of the thesis.
(ii) The topic is rather narrow, and therefore suitable for the dashboard created for illustrative purposes.
(iii) Literature on the topic was familiar to me, because it was used for the study described in Chapter 5; thus, I had sufficient domain knowledge for structuring properties of the research contributions.

(iv) Potential participants of the user evaluation survey were not expected to be familiar with the topic and therefore could focus on evaluating the interface as such without being influenced by their attitude to the topic.

I added relevant papers on the topic to the ORKG knowledge base and created a resource comparison R76906 https://www.orkg.org/orkg/comparison/R76906. The properties of research contributions included datasets, participant countries, methodology, attitudes to ICT, outcome variables, and results (relationships between students' attitudes towards ICT and the outcome variables). The dashboard was designed to present only two properties out of these. Firstly, I decided to visualize the participant countries, that is, the countries of students whose attitudes were explored. Secondly, the relationships were presented between students' attitudes towards ICT and their scores in mathematics, science, and reading.

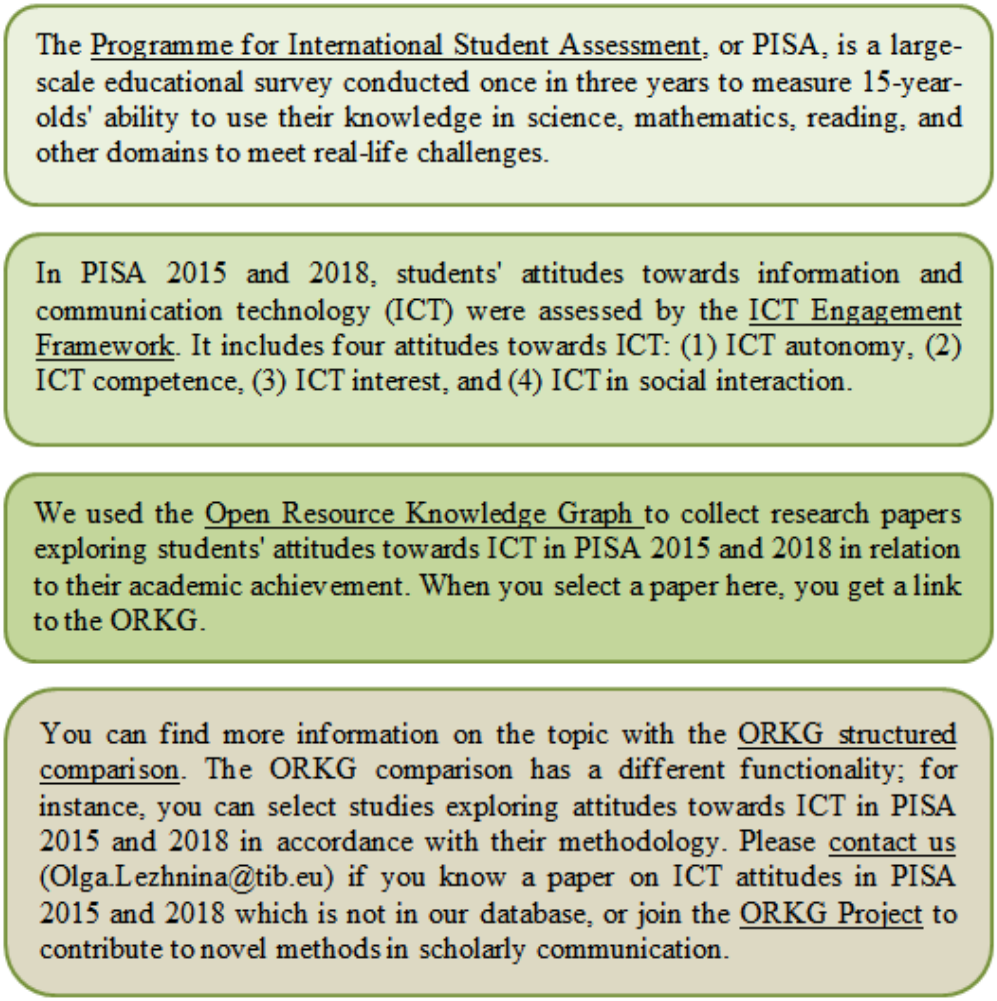

Figure 7.3: **Basic Information for the Dashboard.**

Basic information about PISA, the ICT engagement framework, and the ORKG project was presented to the user (Figure 7.3), and links to the relevant web pages were given. Information presented in the verbal modality was minimized in order to avoid cognitive overload of the user.

The most important aspect, and the benefit, of the dashboard is that it presents information in the visual modality. The requirements for multi-relational dynamic visualisations such as dashboards were elaborated in [38]: they should aim for consistency in selection of their content; schematicity in the formal representation of information; versatility in encoding and setup of visualization; appealingness in graphic design; accessibility of media channel; and effectiveness perceived by the user. Thus, I focused on appeal and ease-of-use of the dashboard from the perspective of the user experience. At the same time, the code was supposed to be easily reproducible for a researcher who intends to run the service locally or create a similar dashboard, which meant that fairly simple visualisations should be used.

For visualizing participant countries, the geographic map of the world was plotted with the pygal library (Python). The user can hover over the map to see the number of studies referring to a specific country. Figure 7.4 shows the results for Finland, which at the moment when the screenshot was taken was included in five studies. The dropdown menu can be used for selecting studies based on the countries of interest. The options included separate countries and lists with more than ten countries in aggregated studies.

Figure 7.4: **The Interactive Map.**

Results of the studies were visualised with the barplots showing the number of studies with a specific finding, that is, a specific relationship between an attitude towards ICT and students' scores in mathematics, reading or science (Figure 7.5).

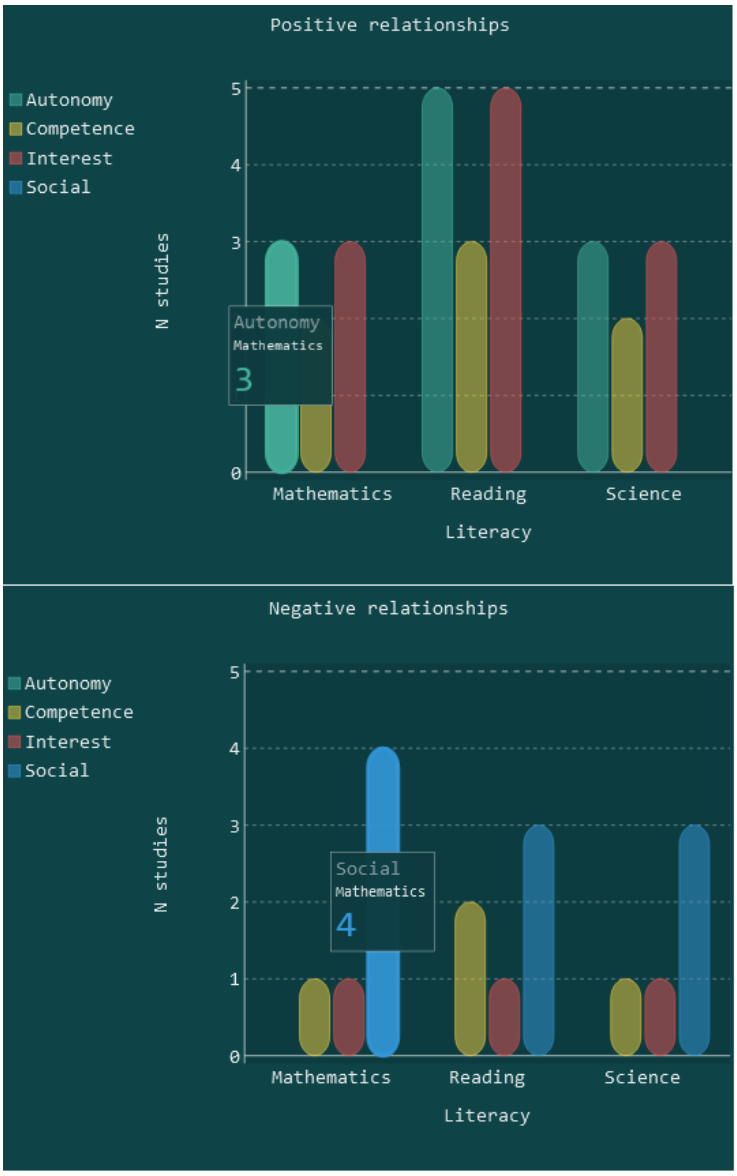

Figure 7.5: **The Interactive Barplots.**

Significance of the relationships was not indicated, as the authors of the studies reported different significance levels. Tooltips could be used to highlight a bar; the user could see which relationship is depicts, and how many studies exploring this relationship could be found in the ORKG database. The users could select studies of their interest with the radio buttons below the plots.

After I developed the dashboard that could be run locally, it was implemented[2] as a use case in the frame of the ORKG research service infrastructure initiative https://www.orkg.org/orkg/usecases/pisa-dashboard/. The evaluation survey was conducted to assess its perception by the users, as described in the next section.

7.2 Dashboard Evaluation

For the dashboard development, my goal was to facilitate versatility in scholarly communication by widening the spectrum of SKG-based interfaces. Therefore, an evaluation was needed to determine whether the ORKG dashboard could be a useful addition to the existing ORKG functionality. The ORKG resource comparison was used as a baseline for the evaluation of the dashboard. It was explicitly stated, though, that both interfaces are integral parts of the ORKG and not mutually exclusive but can complement each other.

7.2.1 User Evaluation Survey

The user evaluation survey consisted of two parts. The participants were asked (1) to evaluate their experiences with the actual services, the ORKG dashboard and the relevant resource comparison; and (2) to assess the potential usefulness of similar services if implemented in their area of research.

In Section A (the first task), I used the short version of the User Experience Questionnaire (the UEQ-S). The instrument was psychometrically validated [112], [224]. It consists of eight pairs of opposite characteristics (confusing/clear, inefficient/efficient, complicated/easy, obstructive/supportive, boring/exciting, not interesting/interesting, conventional/inventive, and usual/leading edge), which the participant evaluates on the scale from -3 to +3. The UEQ-S questions were obligatory to answer.

In Section B (the second task), the participants were asked to evaluate, on the scale from 1 to 5, how advantageous similar services could be for different aspects of scholarly communication if implemented in their area of research. There were five such aspects to be assessed: get acquainted with a new topic; answer a specific question; get an overview of relevant research; explore novel methods of scholarly communication; and make their own research visible for others. The participants also evaluated (on the scale from 1 to 5) the overall usefulness of the dashboard and the resource comparison if jointly implemented in their area of research. Both parts of the survey included open questions, so that the participants could comment on their experience with the dashboard and with the resource comparison separately and reflect on the idea of implementing both services in their area of research.

[2] See the authors' contributions in the related publication.

Finally, as science domains might influence the researchers' attitudes to scholarly communication (see [31], [266]), participants were asked to give relevant information about themselves: whether they worked in technical or humanitarian professions; conduct mostly quantitative or mostly qualitative research; and deal with academic literature rather frequently or only occasionally. Option "other" was included in each of these questions.

7.2.2 Results of Evaluation

The survey was administered via the LimeSurvey service. The participants were invited via social media in professional groups interested in the ORKG and Open Science. The sample (N = 32) included representatives of humanitarian professions (n = 15) and technical professions (n = 13); the participants who chose the option "other" specified their professions as "biology", "nursing", and "art". Mostly quantitative research was conducted by 14 participants, and mostly qualitative by 11 participants. In terms of academic literature, 21 participants dealt with it "rather frequently", and 10 "only occasionally". The scores on the UEQ-S items given to the dashboard and the resource comparison by all participants are presented graphically in Figure 7.6.

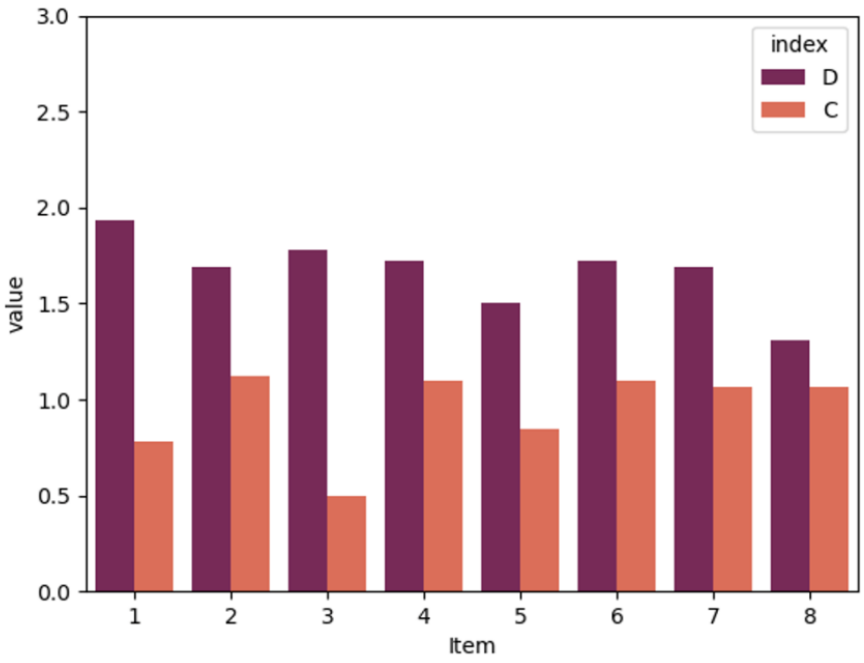

Figure 7.6: **UEQ-S Results.**
D is dashboard, C is comparison. The UEQ-S items: 1) confusing/clear, 2) inefficient/efficient, 3) complicated/easy, 4) obstructive/supportive, 5) boring/exciting, 6) not interesting/interesting, 7) conventional/inventive, and 8) usual/leading edge. The UEQ-S scale starts from -3, only the upper part of the graph is presented for visual clarity.

It can be seen that the dashboard received higher scores on all items than the resource comparison. The difference is especially prominent for item 3, that is, the participants perceived the dashboard as easier to use than the resource comparison.

Subgroups of participants were compared: humanitarian vs technical professions, quantitative vs qualitative research, dealing with academic literature frequently vs occasionally. All subgroups found the dashboard clearer (item 1), easier (item 3), more supportive (item 4), more exciting (item 5), and more interesting (item 6) than the resource comparison.

For participants with technical professions, the dashboard was easier than for those with humanitarian professions. In other items, though, humanitarians gave higher scores to the dashboard than technical professionals (Figure 7.7). Technical professionals considered the dashboard more usual (item 8) than the resource comparison, but it was not the case for humanitarians.

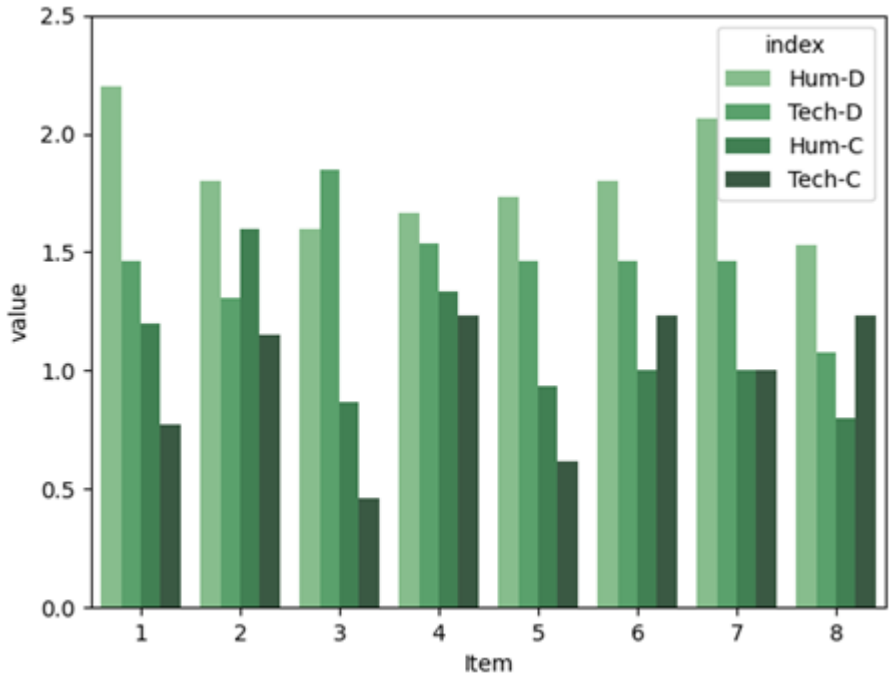

Figure 7.7: **UEQ-S Results, Technical vs Humanitarian Professions.**
D is dashboard, C is comparison. Hum stands for humanitarian professions (n = 15), Tech stands for technical professions (n = 13). The UEQ-S items: 1) confusing/clear, 2) inefficient/efficient, 3) complicated/easy, 4) obstructive/supportive, 5) boring/exciting, 6) not interesting/interesting, 7) conventional/inventive, and 8) usual/leading edge. The UEQ-S scale starts from -3, only the upper part of the graph is presented for visual clarity.

Participants who conducted mostly quantitative research found the dashboard substantially easier, more efficient, and more supportive than those who conducted

mostly qualitative research. The latter group, though, perceived it as more exciting and more inventive than the former (Figure 7.8).

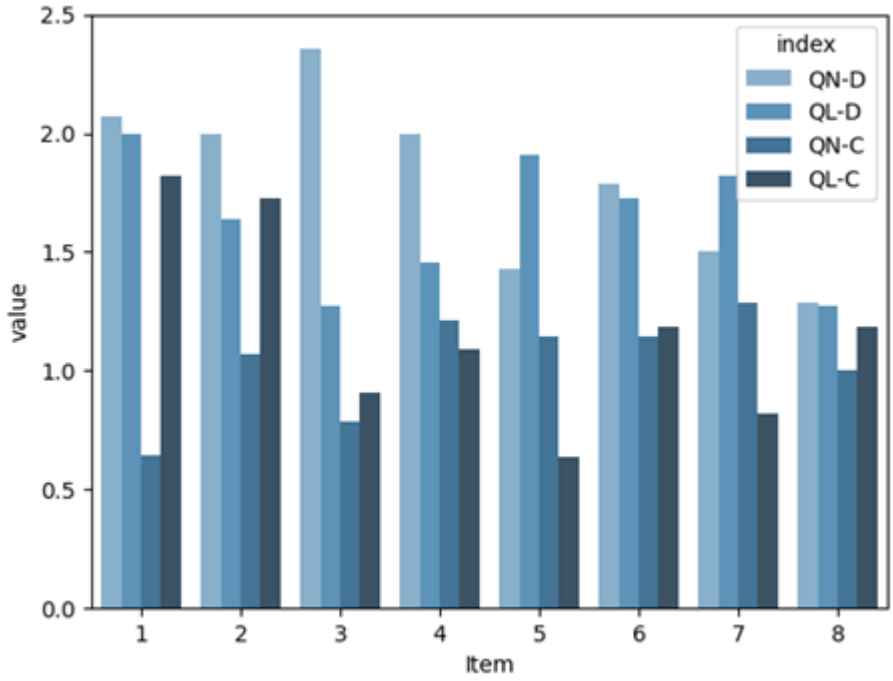

Figure 7.8: **UEQ-S Results, Quantitative vs Qualitative Research.**
D is dashboard, C is comparison. QN stands for "mostly quantitative research" (n = 14), QL stands for "mostly qualitative research" (n = 11). The UEQ-S items: 1) confusing/clear, 2) inefficient/efficient, 3) complicated/easy, 4) obstructive/supportive, 5) boring/exciting, 6) not interesting/interesting, 7) conventional/inventive, and 8) usual/leading edge. The UEQ-S scale starts from -3, only the upper part of the graph is presented for visual clarity.

Participants who dealt with academic literature frequently found the dashboard substantially easier and more exciting than those who dealt with the literature occasionally. The latter group assessed the dashboard as more interesting than the former (Figure 7.9).

When asked to assess similar services if implemented in their area of research, the participants found integration of the dashboard and the resource comparison useful, with the score 4.25 (SD = 0.95) on the scale from 1 to 5. In terms of specific tasks, the respondents were asked to evaluate the usefulness of the services if implemented in their areas of research for the following tasks: (1) get acquainted with a new topic; (2) answer a specific question; (3) get an overview of relevant research; (4) explore novel methods of scholarly communication; and (5) make their own research visible for others. Responses were on the Likert scale from 1 (not helpful at all) to 5 (very helpful).

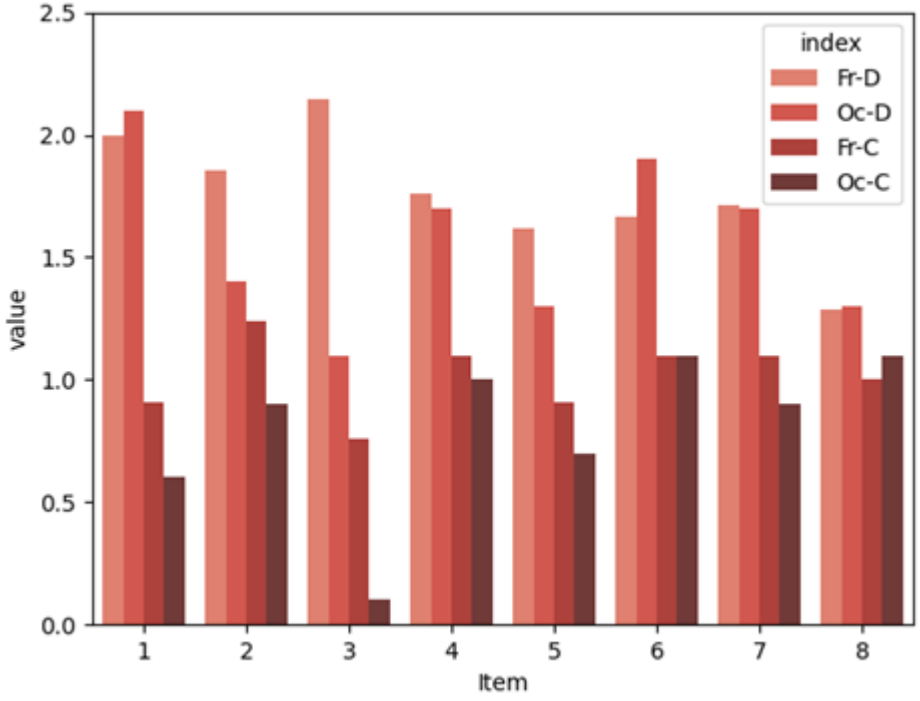

Figure 7.9: **UEQ-S Results, Literature: Frequently vs Occasionally.**
D is dashboard, C is comparison. Fr stands for "rather frequently" (n = 21), Oc stands for "only occasionally" (n = 10). The UEQ-S items: 1) confusing/clear, 2) inefficient/efficient, 3) complicated/easy, 4) obstructive/supportive, 5) boring/exciting, 6) not interesting/interesting, 7) conventional/inventive, and 8) usual/leading edge.

The results are presented in Table 7.1. The dashboard was evaluated as more helpful than the resource comparison for getting acquainted with a new topic, and the resource comparison for answering a specific question. Both services were found useful for getting an overview of relevant research**.**

Table 7.1**: Responses to Section B Items.**

Service	1	2	3	4	5
D	3.87 (1.04)	3.76 (1.15)	4.10 (0.75)	3.77 (1.10)	3.72 (1.25)
R	3.45(1.23)	4.03 (1.00)	4.13 (0.90)	3.60 (1.22)	3.64 (1.13)

D is dashboard; C is comparison. Standard deviations are given in brackets.

In their answers to open questions (the responses were removed from the open access data due to data protection considerations), the participants stressed that both services could be useful for various tasks. The most frequently addressed topic in the comments was ease or difficulty of use of both services. In accordance with the responses to the UEQ-S, some participants called the dashboard easy to use, while

others stated that both services were not very intuitive. Criticism (the direct link to the papers is not provided but hidden two clicks away) and suggestions (highlighting the matching graph when studies are selected) were also aimed at easier use and more coherent presentation of information.

7.3 Summary

In this chapter, I presented implementation and user evaluation of the ORKG dashboard. I developed the dashboard as a multi-relational dynamic visualization tool at the intersection of computer science, graphic design, and human-technology interaction. My aim was to widen the scope of SKG-powered interfaces and explore possible ways of improving the user experience, which would eventually lead to wider acceptance of SKGs by research communities. In terms of method versatility, this approach can be categorized as the range extension: the dashboard was designed as an interface that would supplement the ORKG resource comparison.

The results of the user evaluation survey are preliminary, as the sample was not large enough to give statistical power to inferential tests. In addition, the sampling bias should be taken into account: the sample was not random but consisted of volunteers interested in SKGs or in novel technologies in general. However, these preliminary results might be reassessed with further research on a larger sample, and the current findings are also valuable. The participants perceived the dashboard as easy to use, interesting, and effective. It was considered especially useful for getting acquainted with a new topic, which means that novices in various research areas might benefit from using domain-specific dashboards. The ease-of-use was the most prominent theme in participants' answers to the open questions. In the future, it might be useful to combine inevitable variability of domain-specific dashboards with standardization required for the user familiarity [118], especially in case of domain novices [46]. In the frame of systemic approach [108] adopted in the ORKG, novelty (various dashboards) and familiarity (the resource comparison) can be integrated.

User-friendly interfaces might play a role in facilitating wider acceptance of scholarly knowledge graphs in academic community, which is a prerequisite of the scholarly communication development in the age of digitalisation [96] and research practices appropriate for Open Science [122]. The ORKG dashboard, which I created with the aim to increase versatility in presentation of research results, aids the existing ORKG functionality with the visual modality appreciated by the wider audience. These findings indicate that SKG-powered dashboards might be a valuable addition to other graph-based interfaces in various academic domains.

Chapter 8

Conclusion

Method versatility is crucially important for scientific research, and data science perspectives and approaches that involve a wide range of various methods can be beneficial for other domains, such as social sciences, psychology, and educational science. The aim of this thesis was to facilitate method versatility at measurement, research, and communication stages of the research cycle in relation to human attitudes towards technology. I carried out four studies in relevant areas and introduced method versatility in four different ways as indicated in Figure 8.1.

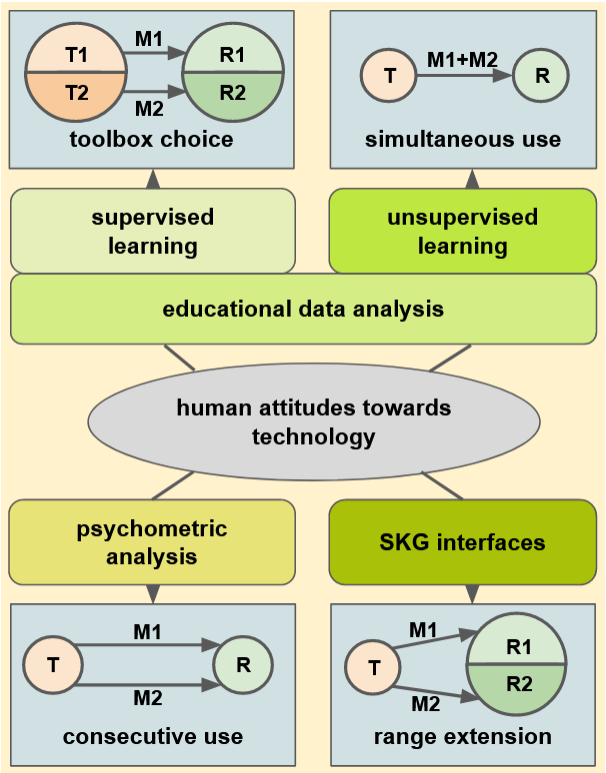

Figure 8.1: **Summary of the Thesis Contributions.**
T stands for task, M for method, R for result.

For psychometric analysis, which is described in Chapter 4, I suggested a strategy with the consecutive use of different methods. The results obtained with CTT methods were reassessed with IRT methods, and the correlation analysis results for construct validity were reassessed with hierarchical clustering of variables.

For supervised learning in educational data analysis in Chapter 5, I implemented a strategy with the toolbox choice of methods: RF as a ML method for missing data imputation and for the classification task, and HLM as a statistical method for the regression task. For unsupervised learning in educational data analysis in Chapter 6, I elaborated an extended strategy for selecting the number of clusters in LCCA with the simultaneous use of model fit, cluster separation, and the stability of partitions criteria.

For SKGs in scholarly communication in Chapter 7, I suggested extending the range of SKG-based interfaces and developed the ORKG-powered dashboard as an interactive visualisation complementing the ORKG resource comparison interface. In this chapter, the impact of these contributions, limitations of the thesis, and ideas for further research are discussed.

8.1 Research Questions Revisited

In the thesis, three RQs were addressed, and four studies were conducted, in which findings on these topics were obtained. This section outlines the impact that these four publications had on the academic community.

8.1.1 RQ1 Contributions

RQ1. How to facilitate method versatility in validation of psychometric instruments related to human attitudes towards technology?

To measure individual aspects of user interaction with technology, valid and reliable instruments are of paramount importance. Hence, psychometricians expect the 2020s to be “the decade of validation” [220], which should be an iterative process conducted via multi-method assessment. My work on the RQ1, as described in Chapter 4, was aimed at increasing method versatility in psychometric analysis of instruments related to human attitudes towards technology.

In my study, a wide range of methods was applied to validation of the ATI scale, a recently developed scale measuring affinity for technology interaction. In terms of method versatility, this was the consecutive use of different methods. Franke, Attig, and Wessel [80], the authors of the scale, used correlation analysis for construct validity, and CTT methods for dimensionality, reliability, and item functioning of the scale. My strategy involved hierarchical clustering of variables for validity, and IRT methods for dimensionality, reliability, and item functioning. For these psychometric

properties, I used the procedures outlined in Dima's study [58] with a number of amendments related to more rigorous assumption testing for CFA and EFA and the choice of R packages for factor retention and multivariate normality test. The strategy can be used at the conceptual level (multimethod psychometric analysis), or as distinct blocks: the CTT-and-IRT procedure (with my amendments) and the validity analysis with hierarchical clustering of variables.

The results of psychometric assessment conducted by the authors of the scale with commonly used methods were overall confirmed, and the ATI was shown to be a valid and reliable unidimensional instrument recommendable for research in the area of human-technology interaction. A potential area for further research and improvement of the scale was indicated, and the findings were shared with the authors of the scale. This relates to a less scalable item that was detected by means of nonparametric IRT (scalability with AISP, the monotonicity test, the invariant item ordering test) and parametric IRT (item fit values); the results were consistent with those obtained by CTT methods (factor loadings and inter-item correlations) but substantially more detailed.

The aspects in which this contribution might be useful to the research community can be summarised as follows:

- the suggested strategy: the consecutive use of different methods
- the application of the existing CTT-and-IRT procedure
- the amendments to the procedure (CFA and EFA assumption checking)
- the use of hierarchical clustering of variables for construct validity analysis
- the results on the ATI scale: reconfirming the previous findings
- the results on the ATI scale: information on the less scalable item
- the open access R code to run the analysis

Since being published, the paper has been cited in a number of psychometric studies (as can be seen in the Article Metrics on the journal website), and the freely available R code has been downloaded by other researchers. The work has been used by researchers selecting a scale to measure users' engagement in technology interaction and psychometricians widening the repertoire of their methods. Multi-method validation procedures will eventually make research in human-technology interaction more replicable, and its results more implementable for enhancing the effectiveness of human-technology interaction.

8.1.2 RQ2 Contributions

RQ2. How to facilitate method versatility in educational research on human attitudes towards technology?

The insufficiently wide spectrum of methods used by educational scientists is a persistent problem in the area. Chapter 5 and Chapter 6 outline the contributions to RQ2 that are related to supervised and unsupervised learning methods in educational research. In both studies, I dealt with data analysis of large-scale educational surveys, PISA and ICILS respectively, and human attitudes towards technology were the topic of research: German students' attitudes to ICT in PISA and German teachers' views on ICT in ICILS.

Chapter 5 describes the contribution in which ML and statistical methods were combined for analysing the PISA data. In terms of method versatility, the suggested strategy was the choice of the most appropriate methods from different toolboxes. I described my analytical choices in completing the three consecutive tasks: (a) missing data imputation; (b) the classification task with proficiency levels (below Level 2, Levels 2–4, or Level 5 and above) as the categorical outcome variable, and (c) the regression task with the literacy score as the continuous outcome variable. For the first task, I selected the RF algorithm based on previous studies showing its effectiveness for this purpose. For the second task, I used another implementation of the RF algorithm, as it was the method of choice due to the ability to handle nonlinearity and interactions. For the third task, however, instead of hierarchical RF or other ML methods, I used HLM from the statistical toolbox, as it was required by characteristics of this specific dataset: for RF, an increase in the group size is more beneficial than an increase in the number of groups, while for HLM the opposite is the case. In conducting HLM, I included plausible values and replicate weights in analysis, which makes this block useful for methodologically rigorous multilevel linear modeling of the PISA data. The results for educational research were reported, and the importance of ICT autonomy for learning was emphasized.

This contribution can be useful to the research community in regard to:

- the suggested strategy: the toolbox choice of different methods
- the analytical decisions regarding RF models and HLM
- RF for missing data imputation: the block for data preprocessing
- RF with model agnostic methods: the block for the classification task
- HLM with plausible values and replicate weights: regression analysis in PISA
- the results for educational researchers: the importance of ICT autonomy
- the open access R code to run the analysis

Chapter 6 describes the contribution that provided a strategy for selecting the number of clusters in LCCA. In terms of method versatility, the strategy was based on the simultaneous use of different methods (in this case, different selection criteria). It was illustrated on the simulated dataset and on the real-world data from ICILS 2018. I showed how the strategy was helpful in selecting generalizable separable clusters in the data, which would not have been possible with the commonly used model fit-based selection. Researchers who conduct LCCA might employ the suggested strategy at the general level (relying on model fit plus cluster separation plus the stability of partitions), or a more specific one (for instance, choosing the BIC elbow heuristic, the minimal BIC, and the minimal ICL for model fit, and the maximal ASW for cluster separation). The R script can be used to conduct end-to-end LCCA**,** or distinct blocks of the code (e.g., the selecting function or preprocessing procedures) can be flexibly employed for the researcher's purposes.

This contribution can be useful to the research community in regard to:

- the suggested strategy: the simultaneous use of different methods
- the data preprocessing block (missing data, visualisations, the decision on dichotomization of the responses)
- the function for cluster selection: results and the plot for the BIC, the ICL, and the ASW
- the function for the stability of partitions: the ARI and the Jaccard index
- the simulated data example
- the end-to-end LCCA example
- the results for educational researchers: generalizable separable clusters in the ICILS data
- the open access R code to run the analysis

Since being published, the first of these papers has been cited in a number of studies, as can be seen in the Article Metrics on the journal website. Most authors citing the paper refer to the methodological aspects of my work, in particular, to the integration of methods from ML and statistical toolboxes. Thus, the message about the importance of method versatility in data analysis of large-scale educational surveys has been heard by the research community. In regard to the LCCA paper that is yet to be cited in other publications, researchers have already used a few blocks of the code (as the author of the code, I have been asked a few questions about it). Data analysis of large-scale educational surveys is an important area of studies for data scientists and educational scientists, and increasing method versatility can be useful for the reproducibility of research findings.

8.1.3 RQ3 Contributions

RQ3. How to facilitate method versatility in communication of research results related to human attitudes towards technology?

SKGs are effective information retrieval tools supporting findability, equal accessibility, and machine readability of academic literature in accordance with open science principles. For their wider acceptance in various research domains, attractive user-friendly interfaces should be developed. Chapter 7 gives an overview of the contribution to the third research question. I developed the SKG-powered dashboard in the frame of the ORKG research service infrastructure initiative. In terms of method versatility, it extends the range of applications of SKGs as effective information retrieval tools which are useful for scholarly communication.

This study contributes to creating and disseminating scholarly knowledge by facilitating the acceptance of SKGs by research communities. Preliminary results of the user evaluation survey showed that the ORKG dashboard is perceived as a more appealing (easier, more interesting, more effective) service than the baseline user interface. The insufficient sample size does not allow conducting inferential tests or making generalizable statements, but the preliminary information can be useful in terms of scores given to the dashboard and the resource comparison by specific groups of users. In addition, answers to open questions about the experience with the dashboard and the resource comparison, which could not be made public due to data privacy considerations, might inform certain aspects of the ORKG development. Further research on a larger sample is needed, but a preliminary conclusion can already be made that the implementation of domain-specific dashboards can support wider acceptance of SKGs, which is crucial for extending the scope of contemporary scholarly communication.

This contribution can be useful to the research community in the following aspects:

- the suggested strategy: the range extension for scholarly communication means
- the dashboard development: technology acceptance principles
- the dashboard development: the open access code (Python, JS, HTML with CSS) to run the service locally
- the evaluation study: instruments and analytical choices
- the evaluation study: the open access code (Python) and the data
- the results of the evaluation: addressing specific audiences
- the ORKG use case: integrating the dashboard and the resource comparison

The specific use case (attitudes towards ICT in PISA) was implemented in the frame of the ORKG research service initiative and can be accessed by the users in combination with the ORKG resource comparison on the topic. New dashboard

prototypes have been developed in the frame of the ORKG[3] since this first dashboard was implemented, and the results of the evaluation reported to the academic community.

8.2 Limitations

Limitations of the thesis are related to (i) the conceptual level in terms of its goals and scope, (ii) methodological strategies suggested, and (iii) the real world datasets used to illustrate these strategies. In this section, I outline these three areas of limitations.

Method versatility is not the final aim but rather a way towards the goal, which is methodologically rigorous scientific research. In this regard, method versatility cannot be sufficient on its own: as I explained in Chapter 1, it is just one of the aspects of the scientific approach. Versatile methods that are not transparently reported, for instance, would not benefit the academic community; therefore, I included transparent reporting in the requirements for my approach. Still, conceptually speaking, focus on versatility bears the risk of overemphasizing one scientific principle at the expense of others, which are similarly important. Another conceptual limitation is the use of categorizations, which are necessary for structuring the work but somewhat artificial, such as the tripartite research cycle, or the four ways of introducing method versatility. In this thesis, I illustrated the application of the consecutive use, the toolbox choice, the simultaneous use, and the range extension in respective studies. However, other approaches to introducing method versatility in data analysis can undoubtedly be found, with various implementations; the scope of my work is not wide enough to explore this topic in more detail.

Methodological strategies of introducing versatility in data analysis that I suggested were selected based not only on their effectiveness but also on their simplicity that would make them comprehensible for a wider audience. This approach imposed some restrictions on the use of the most advanced methodology in studies that I conducted. For instance, in the supervised learning contribution, more sophisticated classification algorithms [125], with model hyperparameters tuned [219], might have performed better in terms of the AUC. In the unsupervised learning contribution, a faster method could be found that would combine the cluster selection function and the stability of partitions function. In the fourth contribution, the dashboard might have been improved in terms of a wider topic and in regard to accessibility (as it is currently available solely on desktop versions of Firefox and Chrome). However, all these amendments would imply making the code less comprehensible for the researchers in such domains as psychology or educational science, who are also a target audience of these studies. In addition, a strategy could cover only a limited number of analytical choices, and therefore, some useful and interesting topics had to be left for further research. For instance, the problem of variable selection could have been dealt with in the supervised

[3] See, for instance, https://doi.org/10.15488/11535

and unsupervised learning contributions, but this topic was beyond the scope of these studies. Last but not least, the Bayesian perspective, which was discussed in literature as an approach to be flexibly combined with the frequentist perspective (see [47], [84], [86]), was not explored in the thesis contributions. The integration of Bayesian and frequentist methods ought to be studied in depth as a distinct topic, which would not have fit the framework of this thesis.

To implement and illustrate the suggested methodological developments, real world datasets were used; the unsupervised learning contribution was partly based on the simulated data. Thus, limitations related to the datasets influenced the methodology of research. In the first contribution, the number of observations in the dataset kindly shared by the authors of the ATI scale was barely sufficient for conducting AISP according to the requirements for this procedure [239], and thus, the results of MSA should be interpreted with caution. The problem of the sample size was even more prominent in the dashboard evaluation study, as the number of participants assessing the services was insufficient for any inferential statistical tests. In the supervised learning and unsupervised learning studies, the sample size was suitable for any methodology that a researcher could possibly apply. However, flaws of sampling design are pervasive even for large-scale educational surveys with strict quality checks of the data. In particular, the following flaws of PISA sampling design were summarized in [114] and [271]: (i) exclusion of students with disabilities from PISA is problematic, as it prevents them from taking part in policy aiming at educational equity; (ii) there is a sampling bias towards slower-maturing students, as some students are excluded because of early graduation, dropout or other forms of attrition; and (iii) age criterion for selecting participants means that students might have had different exposure to curriculum, as some of them might have repeated a class or skipped one. In addition, concerns have been raised about plausible values and their influence on the results [134]. Finally, in all four contributions, the data was used that had been collected with self-report survey items, and thus, the social desirability issue might have biased the responses.

8.3 Further Research

Methodological and practical implications of the thesis can be expanded by further research. The interdisciplinary nature of this work makes it relevant both to data scientists developing novel methods and to domain specialists (educational researchers, psychologists, and social scientists) implementing a wider range of techniques in their respective areas. In the four contributions to the thesis, I gave recommendations on how to facilitate domain-specific research in terms of validating psychometric instruments, flexibly applying a wider spectrum of methods to data analysis of large-scale educational surveys, and developing novel SKG-based interfaces for scholarly communication. Here, I briefly outline more general considerations that can be taken into account by further research on method versatility.

In this thesis, the topic of human attitudes towards technology was selected, which is important for various domains, and out of these, educational data analysis and psychology were specifically paid attention to; in further research, another topic of studies can be chosen. The similar approach of providing a research domain with versatile methods and perspectives from data science can be implemented for medicine, nursing science, and other areas. In addition, ways of introducing method versatility can be different from those explored in this thesis (that is, from the consecutive use, the toolbox choice, the simultaneous use, and the range extension). Moreover, the methods themselves can differ from those scrutinized in this thesis; e.g., the topic of integration of Bayesian and frequentist approaches might be useful and interesting for every science domain. Studying this integration from the data science perspective, such as specifying conditions under which it could be beneficial, is another useful direction of further research.

8.4 Closing Remarks

As the famous saying goes, “all models are wrong” [25], but some can be useful. For finding better approximations of reality, a wide spectrum of flexibly selected methods is required. In this thesis, the problem of method versatility in analysing human attitudes towards technology was approached from the data science perspective in the frame of interdisciplinary effort aimed at providing analytical strategies to other domains. I showed how method versatility can be facilitated at three stages of the research process: validation of psychometric instruments; supervised and unsupervised learning in data analysis of large-scale educational surveys; and scholarly communication via the SKGs with user-friendly interfaces. The versatile methods were introduced as the consecutive use of different techniques, the toolbox choice, the simultaneous use, and the range extension. These and other approaches to method versatility can be explored by further research. Incremental progress in methodology of data analysis and scholarly communication will eventually provide researchers in various domains with the wide spectrum of instruments they need to increase the reproducibility of research findings and deal with the contemporary challenges in data analysis.

Bibliography

[1] Adam, S. P., Alexandropoulos, S.-A. N., Pardalos, P. M., & Vrahatis, M. N. (2019). No Free Lunch Theorem: A Review. In I. C. Demetriou & P. M. Pardalos (Eds.), *Approximation and Optimization* (Vol. 145, pp. 57–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-12767-1_5

[2] Afshartous, D., & de Leeuw, J. (2005). Prediction in multilevel models. *Journal of Educational and Behavioral Statistics, 30*(2), 109–139. https://doi.org/10.3102/10769986030002109

[3] Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management, 39*(6), 1490–1528. https://doi.org/10.1177/0149206313478188

[4] Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access, 7*, 31883–31902. https://doi.org/10.1109/ACCESS.2019.2903568

[5] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

[6] Altman, M., & Cohen, P. N. (2022). The scholarly knowledge ecosystem: Challenges and opportunities for the field of information. *Frontiers in Research Metrics and Analytics, 6*, Article 751553. https://doi.org/10.3389/frma.2021.751553

[7] Anderlucci, L., & Hennig, C. (2014). The clustering of categorical data: A comparison of a model-based and a distance-based approach. *Communications in Statistics - Theory and Methods, 43*(4), 704–721. https://doi.org/10.1080/03610926.2013.806665

[8] Anderson, J. O., Milford, T., & Ross, S. P. (2009). Multilevel modeling with HLM: Taking a second look at PISA. In M. C. Shelley, L. D. Yore, & B. B. Hand (Eds.), *Quality research in literacy and science education: International perspectives and gold standard*s (pp. 263–286). Springer. Netherlands. https://doi.org/10.1007/978-1-4020-8427-0_13

[9] Angioni, S., Salatino, A. A., Osborne, F., Reforgiato, D., & Motta, E. (2020). The AIDA dashboard: Analysing conferences with semantic technologies. *CEUR Workshop Proceedings 2721*, 271–276. http://ceur-ws.org/Vol-2721/

[10] Areepattamannil, S. (2014). International note: What factors are associated with reading, mathematics, and scientific literacy of Indian adolescents? A multilevel examination. *Journal of Adolescence, 37*(4), 367–372. https://doi.org/10.1016/j.adolescence.2014.02.007

[11] Areepattamannil, S., & Santos, I. M. (2019). Adolescent students' perceived information and communication technology (ICT) competence and autonomy:

Examining links to dispositions toward science in 42 countries. *Computers in Human Behavior, 98*, 50–58. https://doi.org/10.1016/j.chb.2019.04.005

[12] Attig, C., Wessel, D., & Franke, T. (2017). Assessing personality differences in human-technology interaction: An overview of key self-report scales to predict successful interaction. In C. Stephanidis (Ed.), HCI International 2017 – Posters' Extended Abstracts (Vol. 713, pp. 19–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-58750-9_3

[13] Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K. E., Vogt, L., Prinz, M., Wiens, V., & Jaradeh, M. Y. (2020). Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis, 44*(3), 516–529. https://doi.org/10.1515/bfp-2020-2042

[14] Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. https://www.ime.unicamp.br/~cnaber/Baker_Book.pdf

[15] Bakker, M., & Wicherts, J.M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PLOS ONE*, *9*(7), 1–9.

[16] Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689–2698. https://doi.org/10.1145/1978942.1979336

[17] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

[18] Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education, 60*(7), 1048–1064. https://doi.org/10.1007/s11162-019-09546-y

[19] Bertoletti, M., Friel, N., & Rastelli, R. (2015). Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *METRON*, *73*, 177–199. https://doi.org/10.1007/s40300-015-0064-5

[20] Bethmann, A., Schierholz, M., Wenzig, K., & Zielonka, M. (2014, September 2–4). *Automatic coding of occupations: Using machine learning algorithms for occupation coding in several German panel surveys* [Conference paper]. World Association for Public Opinion Research (WAPOR) 67th Annual Conference, Nice, France. https://www.researchgate.net/publication/266259591_Automatic_Coding_of_Occupations_Using_Machine_Learning_Algorithms_for_Occupation_Coding_in_Several_German_Panel_Surveys

[21] 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. https://doi.org/10.1109/34.865189

[22] Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018) Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018

[23] Boehmke, B., & Greenwell, B. M. (2020). *Hands-on machine learning with R*. Taylor & Francis. https://bradleyboehmke.github.io/HOML/
[24] Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. *Proceedings of the 2008 SIAM International Conference on Data Mining*, 243–254. https://doi.org/10.1137/1.9781611972788.22
[25] Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949
[26] Brack, A., Hoppe, A., Stocker, M., Auer, S., & Ewerth, R. (2021). Analysing the requirements for an open research knowledge graph: Use cases, quality requirements, and construction strategies. *International Journal on Digital Libraries*, *23*(1), 33–55. https://doi.org/10.1007/s00799-021-00306-x
[27] Branco, P., Ribeiro, R. P., & Torgo, L. (2016). *UBL: An R package for utility-based learning*. arXiv. https://doi.org/10.48550/arXiv.1604.08079
[28] Breiman, L. (2001a). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324
[29] Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726
[30] Brusco, M. J., Shireman, E., & Steinley, D. (2017). A comparison of latent class, K-means, and K-median methods for clustering dichotomous data. *Psychological Methods, 22*(3), 563–580. https://doi.org/10.1037/met0000095
[31] Bu, Y., Lu, W., Wu, Y., Chen, H., & Huang, Y. (2021). How wide is the citation impact of scientific publications? A cross-discipline and large-scale analysis. *Information Processing & Management, 58*(1), Article 102429. https://doi.org/10.1016/j.ipm.2020.102429
[32] Bundsgaard, J., & Gerick, J. (2017). Patterns of students' computer use and relations to their computer and information literacy: Results of a latent class analysis and implications for teaching and learning. *Large-Scale Assessments in Education, 5*(1), 16. https://doi.org/10.1186/s40536-017-0052-8
[33] Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods, 15*(4), 233–234. https://doi.org/10.1038/nmeth.4642
[34] Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 1*19(2), 197-253.
[35] Cao, L. (2018). Data science: A comprehensive overview. *ACM Computing Surveys, 50*(3), 1–42. https://doi.org/10.1145/3076253
[36] Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006
[37] Castro-Alonso, J. C., de Koning, B. B., Fiorella, L., & Paas, F. (2021). Five strategies for optimizing instructional materials: Instructor- and learner-

managed cognitive load. *Educational Psychology Review*. https://doi.org/10.1007/s10648-021-09606-9

[38] Cavaller, V. (2021) Dimensional taxonomy of data visualization: A proposal from communication sciences tackling complexity. *Frontiers in Research Metrics and Analytics, 6,* Article 643533. https://doi.org/10.3389/frma.2021.643533

[39] Cazan, A.-M., & Indreica, S. E. (2014). Need for cognition and approaches to learning among university students. *Procedia - Social and Behavioral Sciences, 127*, 134–138. https://doi.org/10.1016/j.sbspro.2014.03.227

[40] Center for Systems Science and Engineering. (2021, August 3). *COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)*. Johns Hopkins University & Medicine, Coronavirus Resource Center. Retrieved August 3, 2021, from https://coronavirus.jhu.edu/map.html

[41] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development Discussions, 7*(1), 1525–1534. https://doi.org/10.5194/gmdd-7-1525-2014

[42] Chaturvedi, A., Green, P. & Caroll, J. (2001). K-modes clustering. *Journal of Classification, 18*(1), 35–55. https://doi.org/10.1007/s00357-001-0004-3

[43] Chavent, M., Kuentz, V., Liquet, B., & Saracco, L. (2011). *ClustOfVar: An R package for the clustering of variables*. ArXiv. https://doi.org/10.48550/arXiv.1112.0295

[44] Cheah, B. C. (2009, May). Clustering standard errors or modeling multilevel data? *University of Columbia*. https://pdfs.semanticscholar.org/9974/6a51c140030fdc3192695d541db99b644c82.pdf

[45] Çokluk, Ö, Büyüköztürk, S., & Kayri, M. (2010). Analyzing factor structure of the scales by hierarchical clustering analysis: An alternative approach. *Australian Journal of Basic and Applied Sciences, 4*(12), 6397–6403.

[46] Cole, C., Mandelblatt, B., & Stevenson, J. (2002). Visualizing a high recall search strategy output for undergraduates in an exploration stage of researching a term paper. *Information Processing & Management, 38*(1), 37–54. https://doi.org/10.1016/S0306-4573(01)00029-2

[47] Colling, L. J., & Szűcs, D. (2021). Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology, 12*(1), 121–147. https://doi.org/10.1007/s13164-018-0421-4

[48] Coroiu, A. M., Gacenau, R. D., & Pop, H. F. (2016). Discovering patterns in data using ordinal data analysis. *Studia Universitatis Babes-Bolyai, Informatica, 61*(1). https://www.cs.ubbcluj.ro/~studia-i/contents/2016-1/06-CoroiuGaceanuPop.pdf

[49] Cosgrove, J., & Cunningham, R. (2011). A multilevel model of science achievement of Irish students participating in PISA 2006. *The Irish Journal of Education, 39*, 57–73. https://doi.org/10.2307/41548684

[50] Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19*(1), Article 270. https://doi.org/10.1186/s12859-018-2264-5

[51] Currie, A., & Avin, S. (2019). Method pluralism, method mismatch & method bias. *Philosophers' Imprint*, *19*(13), 1–22. http://hdl.handle.net/2027/spo.3521354.0019.013
[52] Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01
[53] Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69–102. https://doi.org/10.3102/0034654308325581
[54] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological), 39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
[55] Denil, M., Matheson, D., & de Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. *Proceedings of the 31st International Conference on Machine Learning, 32*(1), 665–673. http://proceedings.mlr.press/v32/denil14.pdf
[56] DeSantis, S. M., Houseman, E. A., Coull, B. A., Stemmer-Rachamimov, A., & Betensky, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics, 9*(2), 249–262. https://doi.org/10.1093/biostatistics/kxm026
[57] Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The evolution of research on computer-supported collaborative learning. In N. Balacheff et al. (Eds.), *Technology-enhanced learning* (pp. 3–19). Springer Netherlands. https://doi.org/10.1007/978-1-4020-9827-7_1
[58] Dima, A. L. (2018). Scale validation in applied health research: Tutorial for a 6-step R-based psychometrics protocol. *Health Psychology and Behavioral Medicine, 6*(1), 136–161. https://doi.org/10.1080/21642850.2018.1472602
[59] Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734
[60] Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046
[61] Dziak, J. J., Lanza, S. T., & Tan, X. (2014). Effect size, statistical power, and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. Structural Equation Modeling: *A Multidisciplinary Journal, 21*(4), 534–552. https://doi.org/10.1080/10705511.2014.919819
[62] Edison, S. W., & Geissler, G. L. (2003). Measuring attitudes towards general technology: Antecedents, hypotheses and scale development. *Journal of Targeting, Measurement and Analysis for Marketing, 12*(2), 137–156. https://doi.org/10.1057/palgrave.jt.5740104
[63] Eickelmann, B., & Vennemann, M. (2017). Teachers' attitudes and beliefs regarding ICT in teaching and learning in European countries. *European*

Educational Research Journal, *16*(6), 733–761. https://doi.org/10.1177/1474904117725899

[64] Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Exploration Newsletter*, *4*(1), 65–75. https://doi.org/10.1145/568574.568575

[65] Fagginger Auer, M. F., Hickendorff, M., Van Putten, C. M., Béguin, A. A., & Heiser, W. J. (2016). Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students' mathematical strategies. *Applied Measurement in Education, 29*(2), 144–159. https://doi.org/10.1080/08957347.2016.1138959

[66] Faik, I., Barrett, M., & Oborn E. (2020). How information technology matters in societal change: An affordance-based institutional logics perspective. *MIS Quarterly*, *44*(3), 1359–1390. https://doi.org/10.25300/MISQ/2020/14193

[67] Faiola, A. J., Srinivas, P., & Doebbeling, B. N. (2015). A ubiquitous situation-aware data visualization dashboard to reduce ICU clinician cognitive load. *Institute of Electrical and Electronics Engineers*, 439–442. https://doi.org/10.1109/HealthCom.2015.7454540

[68] Fang, Y., & Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis, 56*(3), 468–477. https://doi.org/10.1016/j.csda.2011.09.003

[69] Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: From early developments to recent advancements. *Systems Science & Control Engineering, 2*(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

[70] Few, S. (2004). Dashboard confusion. *Dashboard confusion*. Perceptual Edge. http://www.perceptualedge.com/articles/ie/dashboard_confusion.pdf

[71] Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R. SAGE.

[72] Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

[73] Fletcher, T.D. (2012). *QuantPsyc: Quantitative psychology tools* (R Package Version 1.5) [Computer software]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=QuantPsyc

[74] Flynt, A., & Dean, N. (2016). A survey of popular R packages for cluster analysis. *Journal of Educational and Behavioral Statistics, 41*(2), 205–225. https://eprints.gla.ac.uk/153580/

[75] Fop, M., & Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys, 12*, 18–65. https://doi.org/10.1214/18-SS119

[76] Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA international computer and information literacy study 2018 assessment framework*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-19389-8

[77] Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020a). *Preparing for life in a digital world: IEA International Computer and Information Literacy Study 2018 International Report*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-38781-5

[78] Fraillon, J., Ainley, J., Schulz, W., Duckworth D., & Friedman T. (2020b). *International computer and information literacy study 2018: Technical report*. International Association for the Evaluation of Educational Achievement (IEA). https://www.iea.nl/publications/technical-reports/icils-2018-technical-report
[79] Franke, G., & Sarstedt, M. (2019). Heuristics versus statistics in discriminant validity testing: A comparison of four procedures. *Internet Research, 29*(3), 430–447. https://doi.org/10.1108/IntR-12-2017-0515
[80] Franke, T., Attig, C., & Wessel, D. (2018). A personal resource for technology interaction: Development and validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction, 35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150
[81] Friedman, J. H. (1997, May 14–17). Data mining and statistics: What's the connection? *Computing Science and Statistics* [Symposium]. 29th Symposium on the Interface, Houston, Texas, United States. http://www.stats.org.uk/Friedman1997.pdf
[82] Gabriel, F., Signolet, J., & Westwell, M. (2018). A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. *International Journal of Research & Method in Education, 41*(3), 306–327. https://doi.org/10.1080/1743727X.2017.1301916
[83] Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27*(15), 2865–2873. https://doi.org/10.1002/sim.3107
[84] Gelman, A. (2014). How do we choose our default methods? In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds.). Past, present, and future of statistical science. Chapman and Hall/CRC. https://doi.org/10.1201/b16720
[85] Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*(4), 967–1033. https://doi.org/10.1111/rssa.12276
[86] Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41*(2), 421–440. https://doi.org/10.1177/0149206314547522
[87] Goldhammer, F., Gniewosz, G., & Zylka, J. (2016). ICT engagement in learning environments. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective (pp. 331–351).* Springer Nature. https://doi.org/10.1007/978-3-319-45357-6
[88] Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice, 11*(3), 319–330. https://doi.org/10.1080/0969594042000304618
[89] Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics, 32*(3), 252–286. https://doi.org/10.3102/1076998606298042

[90] Golino, H. F., & Gomes, C. M. A. (2014). Visualizing random forest's prediction results. *Psychology, 5*(19), 2084–2098. https://doi.org/10.4236/psych.2014.519211
[91] Golino, H. F., & Gomes, C. M. A. (2016). Random forest as an imputation method for education and psychology research: Its impact on item fit and difficulty of the Rasch model. *International Journal of Research & Method in Education, 39*(4), 401–421. https://doi.org/10.1080/1743727X.2016.1168798
[92] Golino, H. F., Gomes, C. M. A., & Andrade, D. (2014). Predicting academic achievement of high-school students using machine learning. *Psychology, 5*(18), 2046–2057. https://doi.org/10.4236/psych.2014.518207
[93] Goodboy, A.K., & Kline, R.B. (2017). Statistical and practical concerns with published communication research featuring structural equation modeling. *Communication Research Reports, 34*(1), 68–77.
[94] Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal, 9*(1), 421–436. https://journal.r-project.org/archive/2017/RJ-2017-016/index.html
[95] Grün, B. (2018). *Model-based clustering*. ArXiv. https://doi.org/10.48550/arXiv.1807.01987
[96] Guédon, Guédon, J., Jubb, M., Kramer, B., Laakso, M., Schmidt, B., Simukovic, E., Hansen, J., Kiley, R., Kitson, A., van der Stelt, W., Markram, K., & Patterson, M. (2019). *Future of scholarly publishing and scholarly communication: Report of the Expert Group to the European Commission. Publications Office*. https://doi.org/10.2777/836532
[97] Guillén-Gámez, F. D., & Mayorga-Fernández, M. J. (2020). Identification of variables that predict teachers' attitudes toward ICT in higher education for teaching and research: A study with regression. *Sustainability, 12*(4), Article 1312. https://doi.org/10.3390/su12041312
[98] Güzeller, C. O., & Akin, A. (2014). Relationship between ICT variables and mathematic achievement based on PISA 2006 database: International evidence. *The Turkish Online Journal of Educational Technology, 13*(1), 184–192. http://www.tojet.net/articles/v13i1/13116.pdf https://www.researchgate.net/publication/287464229_Relationship_between_ICT_variables_and_mathematics_achievement_based_on_PISA_2006_database_International_evidence
[99] Hämäläinen, J., Jauhiainen, S., & Kärkkäinen, T. (2017). Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms, 10*(3), 105. https://doi.org/10.3390/a10030105
[100] 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*, Article 2461. https://doi.org/10.3389/fpsyg.2019.02461
[101] Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning, 45*(2), 171–186. https://doi.org/10.1023/A:1010920819831
[102] Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm.

Systems Research and Behavioral Science, 30(2), 194–203. https://doi.org/10.1002/sres.2130

[103] Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modeling and multi-model inference in ecology. *PeerJ, 6*, Article e4794. https://doi.org/10.7717/peerj.4794

[104] Hassenzahl, M., Diefenbach, S., Göritz, A. (2010). Needs, affect, and interactive products – facets of user experience. *Interacting with Computers, 22* (5), 353–362. https://doi.org/10.1016/j.intcom.2010.04.002

[105] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[106] Hayward, E. (2022, October 5). What is a data dashboard? Retrieved from https://www.klipfolio.com/blog/what-is-a-data-dashboard

[107] Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection - a review and recommendations for the practicing statistician. *Biometrical Journal, 60*(3), 431–449. https://doi.org/10.1002/bimj.201700067

[108] Helkkula, A., Kowalkowski, C., & Tronvoll, B. (2018). Archetypes of service innovation: Implications for value cocreation. *Journal of Service Research, 21*(3), 284–301. https://doi.org/10.1177/1094670517746776

[109] Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis, 52*, 258–271. https://doi.org/10.1016/j.csda.2006.11.025

[110] Hennig, C. (2015). *Clustering strategy and method selection*. ArXiv. https://doi.org/10.48550/arXiv.1503.02059

[111] Hennig, C., & Liao, T. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*(3), 309–369. https://doi.org/10.1111/j.1467-9876.2012.01066.x

[112] Hinderks, A., Schrepp, M., & Thomaschewski, J. (2018). A benchmark for the short version of the user experience questionnaire. *Proceedings of the 14th International Conference on Web Information Systems and Technologies*, 373–377. https://doi.org/10.5220/0007188303730377

[113] Hogan, T. P., & Agnello, J. (2004). An Empirical Study of Reporting Practices Concerning Measurement Validity. *Educational and Psychological Measurement, 64*(5), 802–812. https://doi.org/10.1177/0013164404264120

[114] Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

[115] Hornbæk, K., & Hertzum, M. (2017). Technology acceptance and user experience: A review of the experiential component in HCI. *ACM Transactions on Computer-Human Interaction, 24*(5), 1–30. https://doi.org/10.1145/3127358

[116] Howard, M.C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664

[117] Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

[118] Hu, P. J.-H., Hu, H., & Fang, X. (2017). Examining the mediating roles of cognitive load and performance outcomes in user satisfaction with a website: A field quasi-experiment. *MIS Quarterly*, *41*(3), 975–987. https://doi.org/10.25300/MISQ/2017/41.3.14

[119] Hu, X., Gong, Y., Lai, C., & Leung, F. K. S. (2018). The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis. *Computers & Education, 125*, 1–13. https://doi.org/10.1016/j.compedu.2018.05.021

[120] Huang, H.-M., Rauch, U., & Liaw, S.-S. (2010). Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach. *Computers & Education*, *55*(3), 1171–1182. https://doi.org/10.1016/j.compedu.2010.05.014

[121] Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 335–348. https://doi.org/10.1145/3514094.3534196

[122] Ignat, T., Ayris, P., Gini, B., Stepankova, O., Özdemir, D., Bal, D., & Deyanova, Y. (2021). Perspectives on open science and the future of scholarly communication: Internet trackers and algorithmic persuasion. Frontiers in Research Metrics and Analytics, *6*, Article 748095. https://doi.org/10.3389/frma.2021.748095

[123] International Association for the Evaluation of Educational Achievement. (n.d.). *International computer and information literacy study: Data repository: ICISL 2018: SPSS data & documentation* [Data set]. https://www.iea.nl/data-tools/repository/icils

[124] Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23. https://doi.org/10.1037/a0014694

[125] Janitza, S., Tutz, G., & Boulesteix, A.-L. (2014, December). *Random forests for ordinal response data: Prediction and variable selection* (Technical Report No. 174). University of Munich, Department of Statistics. https://epub.ub.uni-muenchen.de/22003/

[126] Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., & Auer, S. (2019). Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. *Proceedings of the*

10th International Conference on Knowledge Capture, 243–246. https://doi.org/10.1145/3360901.3364435

[127] Joenssen, D.W. & Vogel, J. (2012). A power study of goodness-of-fit tests for multivariate normality implemented in R. *Journal of Statistical Computation and Simulation*, *84*(5), 1055–1078. https://doi.org/10.1080/00949655.2012.739620

[128] Kamata, A., Kara, Y., Patarapichayatham, C., & Lan, P. (2018). Evaluation of analysis approaches for latent class analysis with auxiliary linear growth model. *Frontiers in Psychology*, *9*, Article 130. https://doi.org/10.3389/fpsyg.2018.00130

[129] Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen – Der Fragebogen TA-EG [Measuring affinity to technology – – the questionnaire TA-EG]. In A. Lichtenstein, C. Stößel, & C. Clemens (Eds.), *Der Mensch im Mittelpunkt technischer Systeme: 8. Berliner Werkstatt Mensch-Maschine-Systeme: 7. bis 9. Oktober 2009* (pp. 196–201). VDI-Verlag.

[130] Kaufman, L., & Rousseeuw, P. J. (1990) *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470316801

[131] Kenett, R. S. (2015). Statistics: A life cycle view. *Quality Engineering*, *27*(1), 111–121. https://doi.org/10.1080/08982112.2015.968054

[132] Kleiman, R. S., & Page, D. (2019). AUCμ: A performance metric for multi-class machine learning models. *Proceedings of the 36th International Conference on Machine Learning*, *97*, 3439–3447. http://proceedings.mlr.press/v97/kleiman19a/kleiman19a.pdf

[133] Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software, 74*(7). https://doi.org/10.18637/jss.v074.i07

[134] Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231. https://doi.org/10.1007/s11336-013-9347-z

[135] Kunina-Habenicht, O., & Goldhammer, F. (2020). ICT engagement: A new construct and its assessment in PISA 2015. Large-Scale Assessments in *Education*, *8*(1), Article 6. https://doi.org/10.1186/s40536-020-00084-z

[136] Kurteva, A., & De Ribaupierre, H. (2021). Interface to query and visualise definitions from a knowledge base. In M. Brambilla, R. Chbeir, F. Frasincar, & I. Manolescu (Eds.), *Web engineering* (Vol. 12706, pp. 3–10). Springer International Publishing. https://doi.org/10.1007/978-3-030-74296-6_1

[137] Lah, U., Lewis, J. R., & Šumak, B. (2020). Perceived usability and the modified technology acceptance model. *International Journal of Human-Computer Interaction*, *36*(13), 1216–1230. https://doi.org/10.1080/10447318.2020.1727262

[138] LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17*(4), 433–451. https://doi.org/10.1177/1094428114541701

[139] Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., . . . Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour, 2*, 168–171. https://doi.org/10.1038/s41562-018-0311-x

[140] Lanza, S. T., Bray, B. C., & Collins, L. M. (2012). An introduction to latent class and latent transition analysis. In I. Weiner, J. A. Schinka & W. F. Velicer (Eds.), *Handbook of Psychology* (2nd ed., Vol. 2, pp. 691–716). John Wiley & Sons.

[141] Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (Vol. 4, pp. 362–472). Princeton University Press.

[142] Lee, J., Jung, K., & Park, J. (2020). Detecting conditional dependence using flexible Bayesian latent class analysis. *Frontiers in Psychology, 11*, Article 1987. https://doi.org/10.3389/fpsyg.2020.01987

[143] Lent, H., Hahn-Powell, G., Haug-Baltzell, A., Davey, S., Surdeanu, M., & Lyons, E. (2018) Science Citation Knowledge Extractor. *Frontiers in Research Metrics and Analytics 3*(35). https://doi.org/10.3389/frma.2018.00035

[144] Lewis, J.R. (2015) Introduction to the special issue on usability and user experience: Psychometrics. *International Journal of Human-Computer Interaction*, *31*(8), 481–483. https://doi.org/10.1080/03610918.2015.1064643

[145] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News 2*(3), 18–22. https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

[146] Lin, S.-W., & Tai, W.-C. (2015). Latent class analysis of students' mathematics learning strategies and the relationship between learning strategy and mathematical literacy. *Universal Journal of Educational Research, 3*(6), 390–395. https://doi.org/10.13189/ujer.2015.030606

[147] Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6*(8). https://doi.org/10.1186/s40536-018-0061-2

[148] Lorenceau, A., Marec, C., & Mostafa, T. (2019). *Upgrading the ICT questionnaire items in PISA 2021* (OECD Education Working Paper No. 202). Organisation for Economic Co-operation and Development. https://doi.org/10.1787/d0f94dc7-en

[149] Lu, Y., & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-Scale Assessments in Education, 3*(1). https://doi.org/10.1186/s40536-015-0012-0

[150] Lüdecke, D. (2019). *sjPlot: Data visualization for statistics in social science* (R Package Version 2.7.2) [Computer software]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=sjPlot

[151] Lüdecke, D., Makowski, D., Waggoner, P., & Patil, I. (2020). *performance: Assessment of regression models performance* (R Package Version 0.4.6)

[Computer software]. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/performance/index.html
[152] Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*, 1494–1502. https://doi.org/10.3758/s13428-016-0809-y
[153] Luu, K., & Freeman, J. G. (2011). An analysis of the relationship between information and communication technology (ICT) and scientific literacy in Canada and Australia. *Computers & Education, 56(*4), 1072–1082. https://doi.org/10.1016/j.compedu.2010.11.008
[154] MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19
[155] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281–297. https://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992
[156] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE, 13*(3), Article e0194889. https://doi.org/10.1371/journal.pone.0194889
[157] Manovich, L. (2011). What is visualization? *Visual Studies, 26*(1), 36–49. https://doi.org/10.1080/1472586X.2011.548488
[158] Marbac, M., & Sedki, M. (2019). VarSelLCM: An R/C++ package for variable selection in model-based clustering of mixed-data with missing values. *Bioinformatics, 35*(7), 1255–1257. https://doi.org/10.1093/bioinformatics/bty786
[159] Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Statistical analysis* (Vol. 2, pp. 551–611). Oxford University Press.
[160] McCoach, D. B. (2010). Dealing with dependence (Part II): A gentle introduction to hierarchical linear modeling. *Gifted Child Quarterly, 54*(3), 252–256. https://doi.org/10.1177/0016986210373475
[161] Meier, S., Gebel-Sauer, B., & Schubert, P. (2021). Knowledge graph for the visualisation of CRM objects in a Social Network of Business Objects (SoNBO): Development of the SoNBO Visualiser. *Procedia Computer Science, 181*, 448–456. https://doi.org/10.1016/j.procs.2021.01.190
[162] Meng, L., Qiu, C., & Boyd-Wilson, B. (2019). Measurement invariance of the ICT engagement construct and its association with students' performance in China and Germany: Evidence from PISA 2015 data. *British Journal of Educational Technology, 50*(6), 3233–3251. https://doi.org/10.1111/bjet.12729
[163] Misztal, M. A. (2019). Comparison of selected multiple imputation methods for continuous variables – preliminary simulation study results. . *Acta Universitatis Lodziensis. Folia Oeconomica, 6*(339), 73–98. https://doi.org/10.18778/0208-6018.339.05

[164] Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

[165] Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: Impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology, 2*, Article 74. https://doi.org/10.3389/fpsyg.2011.00074

[166] Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*, Article 20170213. https://doi.org/10.1098/rsif.2017.0213

[167] Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

[168] Nalbantoğlu-Yılmaz, F. (2019). Comparison of different estimation methods used in confirmatory factor analyses in non-normal data: A Monte Carlo study. *International Online Journal of Educational Sciences, 11*(4), 131–140. https://doi.org/10.15345/iojes.2019.04.010

[169] Nationales Bildungspanel. (2019). Forschungsprojekte mit NEPS-Daten [Research projects with NEPS data]. Retrieved from https://www.neps-data.de/de-de/datenzentrum/forschungsprojekte.aspx

[170] Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535–569. https://doi.org/10.1080/10705510701575396

[171] Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science, 4*(4), 440–461. https://doi.org/10.1037/tps0000176

[172] O'Connell, A. A., & Liu, X. (2011). Model diagnostics for proportional and partial proportional odds models. *Journal of Modern Applied Statistical Methods, 10*(1), Article 15. https://doi.org/10.22237/jmasm/1304223240

[173] Odell, B., Galovan, A. M., & Cutumisu, M. (2020). The relation between ICT and science in PISA 2015 for Bulgarian and Finnish students. *EURASIA Journal of Mathematics, Science and Technology Education, 16*(6). https://doi.org/10.29333/ejmste/7805

[174] Oelen, A., Jaradeh, M. Y., Farfar, K. E., Stocker, M., & Auer, S. (2019). Comparing research contributions in a scholarly knowledge graph. *CEUR Workshop Proceedings, 2526*, 21–26. https://doi.org/10.15488/9388

[175] Oelen, A., Jaradeh, M. Y., Stocker, M., & Auer, S. (2020). Generate FAIR literature surveys with scholarly knowledge graphs. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 97–106. https://doi.org/10.1145/3383583.3398520

[176] Organisation for Economic Co-operation and Development. (n.d.-a). *PISA 2015 database* [Data set]. Retrieved from http://www.oecd.org/pisa/data/2015database/
[177] Organisation for Economic Co-operation and Development. (n.d.-b). *PISA 2018 database* [Data set]. Retrieved from http://www.oecd.org/pisa/data/2018database/
[178] Organisation for Economic Co-operation and Development. (2009). *PISA Data Analysis Manual: SPSS* (2nd ed.). https://doi.org/10.1787/9789264056275-en
[179] Organisation for Economic Co-operation and Development. (2016a). *Programme for international student assessment (PISA): Results from PISA 2015* [Country note for Germany]. https://www.oecd.org/pisa/PISA-2015-Germany.pdf
[180] Organisation for Economic Co-operation and Development. (2016b). *PISA: Low-performing students: Why they fall behind and how to help them succeed.* https://doi.org/10.1787/9789264250246-en
[181] Organisation for Economic Co-operation and Development. (2017a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving.* https://doi.org/10.1787/9789264281820-en
[182] Organisation for Economic Co-operation and Development. (2017b). *PISA 2015 technical report*. http://www.oecd.org/pisa/data/2015-technical-report/
[183] Organisation for Economic Co-operation and Development. (2019a). *PISA 2018 assessment and analytical framework*. https://doi.org/10.1787/b25efab8-en
[184] Organisation for Economic Co-operation and Development. (2019b). *PISA 2018: Insights and interpretations.* https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf
[185] Organisation for Economic Co-operation and Development. (2019c). *PISA 2018 results (Volume 1): What students know and can do.* https://doi.org/10.1787/5f07c754-en
[186] Organisation for Economic Co-operation and Development. (2019d). *PISA 2021 ICT framework*. https://www.oecd.org/pisa/sitedocument/PISA-2021-ICT-framework.pdf
[187] Ortiz Vivar, J., Segarra, J., Villazón-Terrazas, B., & Saquicela, V. (2020). REDI: Towards knowledge graph-powered scholarly information management and research networking. *Journal of Information Science 48*(2), 1–15. https://doi.org/10.1177/0165551520944351
[188] Pauwels, K., Ambler, T., Clark, B. H., LaPointe, P., Reibstein, D., Skiera, B., Wierenga, B., & Wiesel, T. (2009). Dashboards as a service: Why, what, how, and what research is needed? *Journal of Service Research, 12*(2), 175–189. https://doi.org/10.1177/1094670509344213
[189] Pérez-Ortega, J., Almanza-Ortega, N. N., Vega-Villalobos, A., Pazos-Rangel, R. A., Zavala-Díaz, C., & Martínez-Rebollar, A. (2020). The K-means

algorithm evolution. In K. Sud, P. Erdogmus, & S. Kadry (Eds.), *Introduction to data science and machine learning* (pp. 69–90). IntechOpen. https://doi.org/10.5772/intechopen.77469

[190] Petersen, K. J., Qualter, P., & Humphrey, N. (2019). The application of latent class analysis for investigating population child mental health: A systematic review. *Frontiers in Psychology, 10*, Article 1214. https://doi.org/10.3389/fpsyg.2019.01214

[191] Petko, D., Cantieni, A., & Prasse, D. (2017). Perceived quality of educational technology matters: A secondary analysis of students' ICT use, ICT-related attitudes, and PISA 2012 test scores. *Journal of Educational Computing Research, 54*(8), 1070–1091. https://doi.org/10.1177/0735633116649373

[192] Plumbley, M. D., & Abdallah, S. A. (2006). Information theory and sensory perception. In J. A. Bryant, M. A. Atherton, & M. W. Collins (Eds.), *Design and information in biology: From molecules to systems* (Vol. 27, pp. 205–233). WIT Press. https://doi.org/10.2495/978-1-85312-853-0/07

[193] Pokropek, A., Costa, P., Flisi, S., & Biagi, F. (2018). *Low achievers, teaching practices and learning environment* (Joint Research Centre Technical Report EUR 29387 EN). Publications Office of the European Union. https://doi.org/10.2760/973882

[194] Qin, Y., Cao, H., & Xue, L. (2020). Research and application of knowledge graph in teaching: Take the database course as an example. *Journal of Physics: Conference Series, 1607*(1), Article 012127. https://doi.org/10.1088/1742-6596/1607/1/012127

[195] Qiu, D., & Malthouse, E. (2009). Cluster analysis with general latent class model. In J. Wang (Ed.), *Encyclopedia of data warehousing and mining* (2nd ed., pp. 225–230). IGI Global. https://doi.org/10.4018/978-1-60566-010-3.ch037

[196] R Core Team. (2022). R: *A language and environment for statistical computing* (R Version 4.0.2) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

[197] Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 805–827. https://doi.org/10.1111/j.1467-985X.2006.00426.x

[198] Raiche, G. (2022). *nFactors: Parallel analysis and other non graphical solutions to the Cattell scree test* (R Package Version 2.3.3.) [Computer software]. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/nFactors/index.html

[199] Ranalli, M., & Rocci, R. (2015). Clustering methods for ordinal data: A comparison between standard and new approaches. In I. Morlini, T. Minerva, & M. Vichi. (Eds.), *Advances in statistical models for data analysis* (pp. 221–229). Springer International Publishing. https://doi.org/10.1007/978-3-319-17377-1

[200] Reinanda, R., Meij, E., & de Rijke, M. (2020). Knowledge graphs: An information retrieval perspective. *Foundations and Trends in Information Retrieval, 14*(4), 289–444. https://doi.org/10.1561/1500000063

[201] Revelle, W. (1978). ICLUST: A cluster analytic approach to exploratory and confirmatory scale construction. *Behavior Research Methods & Instrumentation, 10*(5), 739–742. https://doi.org/10.3758/BF03205389

[202] Revelle, W. (2018) *psych: Procedures for personality and psychological research*. (R Package Version 1.8.12) [Computer software]. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/psych/index.html

[203] Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E., & Köller, O. (Eds.). (2016). *PISA 2015: Eine Studie zwischen Kontinuität und Innovation* [PISA 2015: A study between continuity and innovation]. Waxmann. https://www.waxmann.com/index.php?eID=download&buchnr=3555

[204] Rienties, B., Køhler Simonsen, H., & Herodotou, C. (2020). Defining the boundaries between artificial intelligence in education, computer-supported collaborative learning, educational data mining, and learning analytics: A need for coherence. *Frontiers in Education, 5*, Article 128. https://doi.org/10.3389/feduc.2020.00128

[205] Rivera, P. M., Fincham, F. D., & Bray, B. C. (2018). Latent classes of maltreatment: A systematic review and critique. *Child Maltreatment, 23*(1), 3–24. https://doi.org/10.1177/1077559517728125

[206] Rizun, M. (2019). Knowledge graph application in education: A literature review. *Acta Universitatis Lodziensis. Folia Oeconomica*, *3*(342), 7–19. https://doi.org/10.18778/0208-6018.342.01

[207] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1). https://doi.org/10.1186/1471-2105-12-77

[208] Rodrigues, M., & Biagi, F. (2017). Digital technologies and learning outcomes of students from low socio-economic background: An analysis of PISA 2015 (Joint Research Centre Technical Report EUR 28688 EN). Publications Office of the European Union. https://doi.org/10.2760/415251

[209] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., & Amancio, D. R. (2019). Clustering algorithms: A comparative approach. *PLOS ONE, 14*(1), Article e0210236. https://doi.org/10.1371/journal.pone.0210236

[210] Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*(1), 6–14. https://doi.org/10.1177/1094428120968614

[211] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery, 10*(3), Article e1355. https://doi.org/10.1002/widm.1355

[212] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[213] Ryo, M., & Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere, 8*(11), Article e01976. https://doi.org/10.1002/ecs2.1976

[214] Saarela, M., Yener, B., Zaki, M. J., & Kärkkäinen, T. (2016). Predicting math performance from raw large-scale educational assessments data: A machine learning approach. *JMLR Workshop and Conference Proceedings: MLDEAS Workshop Papers of the 33rd International Conference on Machine Learning*, *48*. http://medianetlab.ee.ucla.edu/papers/ICMLWS3.pdf

[215] Sabou, M., Simperl, E., Blomqvist, E., Groth, P., Kirrane, S., de Melo, G., Mons, B., Paulheim, H., Pintscher, L., Presutti, V., Sequeda, J. F., & Shimizu, C. M. (2018). 3.24: Human and social factors in knowledge graphs [Short talk overview]. *Dagstahl Reports*, *9*(8), 100–104. https://doi.org/10.4230/DAGREP.8.9.29

[216] Santos, H., Dantas, V., Furtado, V., Pinheiro, P., & McGuinness, D. L. (2017). From data to city indicators: A knowledge graph for supporting automatic generation of dashboards. In E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, & O. Hartig (Eds.), *The semantic web* (Vol. 10250, pp. 94–108). Springer International Publishing. https://doi.org/10.1007/978-3-319-58451-5_7

[217] Sarrafzadeh, B., Vtyurina, A., Lank, E., & Vechtomova, O. (2016). Knowledge graphs versus hierarchies: An analysis of user behaviours and perspectives in information seeking. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, 91–100. https://doi.org/10.1145/2854946.2854958

[218] Say, A., Fathalla, S., Vahdati, S., Lehmann, J., & Auer, S. (2020). Semantic representation of physics research data: *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 64–75. https://doi.org/10.5220/0010111000640075

[219] Schiffner, J., Bischl, B., Lang, M., Richter, J., Jones, Z. M., Probst, P., Pfisterer, F., Gallo, M., Kirchhoff, D., Kühn, T., Thomas, J., & Kotthoff, L. (2016, September 18). *mlr tutorial*. ArXiv. https://doi.org/10.48550/arXiv.1609.06146

[220] Schimmack, U. (2019). *The Validation Crisis in Psychology* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/q247m

[221] Schmettow, M., Noordzij, M. L., & Mundt, M. (2013). An implicit test of UX: Individuals differ in what they associate with computers. *Proceedings of the Extended Abstracts on Human Factors in Computing Systems*, 2039–2048. https://doi.org/10.1145/2468356.2468722

[222] Schmiege, S. J., Masyn, K. E., & Bryan, A. D. (2018). Confirmatory latent class analysis: Illustrations of empirically driven and theoretically driven model constraints. *Organizational Research Methods*, *21*(4), 983–1001. https://doi.org/10.1177/1094428117747689

[223] Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304–321. https://doi.org/10.1177/0734282911406653

[224] Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(6), 103–108. https://doi.org/10.9781/ijimai.2017.09.001

[225] Schwab, C. G. G., Dichter, M. N., & Berwig, M. (2018). Item distribution, internal consistency, and structural validity of the German version of the DEMQOL and DEMQOL–proxy. *BMC Geriatrics, 18*(1). https://doi.org/10.1186/s12877-018-0930-0

[226] Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://projecteuclid.org/download/pdf_1/euclid.aos/1176344136

[227] Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology, 179*(6), 764–774. https://doi.org/10.1093/aje/kwt312

[228] Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods, 49*(1), 282–293. https://doi.org/10.3758/s13428-015-0697-6

[229] Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

[230] Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., . . . Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. https://doi.org/10.1177/2515245917747646

[231] Singh, J. (2004). Tackling measurement problems with item response theory. *Journal of Business Research, 57*(2), 184–208. https://doi.org/10.1016/S0148-2963(01)00302-2

[232] Soroya, S. H., Farooq, A., Mahmood, K., Isoaho, J., & Zara, S. (2021). From information seeking to information avoidance: Understanding the health information behavior during a global health crisis. *Information Processing & Management, 58*(2), Article 102440. https://doi.org/10.1016/j.ipm.2020.102440

[233] Spoon, K., Beemer, J., Whitmer, J. C., Fan, J., Frazee, J. P., Stronach, J., Bohonak, A. J., & Levine, R. A. (2016). Random forests for evaluating pedagogy and informing personalized learning. *Journal of Educational Data Mining, 8*(2), 20–50. https://doi.org/10.5281/zenodo.3554595

[234] Stahl, D., & Sallis, H. (2012). Model-based cluster analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 4*(4), 341–358. https://doi.org/10.1002/wics.1204

[235] Stamovlasis, D., Papageorgiou, G., Tsitsipis, G., Tsikalas, T., & Vaiopoulou, J. (2018). Illustration of step-wise latent class modeling with covariates and taxometric analysis in research probing children's mental models in learning sciences. *Frontiers in Psychology, 9*, Article 532. https://doi.org/10.3389/fpsyg.2018.00532

[236] Steenwinckel, B., Vandewiele, G., Rausch, I., Heyvaert, P., Taelman, R., Colpaert, P., Simoens, P., Dimou, A., De Turck, F., & Ongenae, F. (2020).

Facilitating the analysis of COVID-19 literature through a knowledge graph. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), *The semantic web* (Vol. 12507, pp. 344–357). Springer International Publishing. https://doi.org/10.1007/978-3-030-62466-8_22

[237] Stekhoven, D. J. (2013). *missForest: Nonparametric missing value imputation using random forest* (R Package Version 1.4) [Computer software]. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/missForest/index.html

[238] Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken Scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology, 12*(1). https://doi.org/10.1186/1471-2288-12-74

[239] Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken Scale analysis. *Educational and Psychological Measurement*, *74*(5), 809–822. https://doi.org/10.1177/0013164414529793

[240] Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources, and a solution. *BMC Bioinformatics, 8*, Article 25. https://doi.org/10.1186/1471-2105-8-25

[241] Sully, J. (1882). Versatility. *Mind*, *7*(27), 366–380. http://www.jstor.org/stable/2246716

[242] Sun, L., Bradley, K. D., & Akers, K. (2012). A multilevel modeling approach to investigating factors impacting science achievement for secondary school students: PISA Hong Kong sample. *International Journal of Science Education, 34*(14), 2107–2125. https://doi.org/10.1080/09500693.2012.708063

[243] Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

[244] Tam, K. Y. Y., van Tilburg, W. A. P., Chan, C. S., Igou, E. R., & Lau, H. (2021). Attention drifting in and out: The boredom feedback model. *Personality and Social Psychology Review*, 25(3), 251–272. https://doi.org/10.1177/10888683211010297

[245] Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 363–377. https://doi.org/10.1002/sam.11348

[246] Turki, H., Hadj Taieb, M. A., Ben Aouicha, M., Fraumann, G., Hauschke, C., & Heller, L. (2021). Enhancing knowledge graph extraction and validation from scholarly publications using bibliographic metadata. *Frontiers in Research Metrics and Analytics*, *6*, Article 694307. https://doi.org/10.3389/frma.2021.694307

[247] Vargas, H., Buil Aranda, C., & Hogan, A. (2019). RDF explorer: A visual query builder for semantic web knowledge graphs [Demonstration]. *CEUR Workshop Proceedings*, *2456*, 229–232. http://ceur-ws.org/Vol-2456/

[248] Vargas-Quesada, B., Chinchilla-Rodríguez, Z., & Rodriguez, N. (2017). Identification and visualization of the intellectual structure in graphene

research. *Frontiers in Research Metrics and Analytics, 2*. https://doi.org/10.3389/frma.2017.00007

[249] Venkatesh, V., Thong, J., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the Association for Information Systems, 17*(5), 328–376. https://doi.org/10.17705/1jais.00428

[250] Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

[251] Vermunt, J. K., & Magidson, J. (2004). Local independence. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The Sage encyclopedia of social sciences research methods* (pp. 580–581). Sage.

[252] Vogt, L., D'Souza, J., Stocker, M., & Auer, S. (2020). Toward representing research contributions in scholarly knowledge graphs using knowledge graph cells. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 107–116. https://doi.org/10.1145/3383583.3398530

[253] Wagner, S., & Wagner, D. (2007). *Comparing clusterings – an overview* (Technical Report No. 2006-04). Karlsruher Institute of Technology, Department of Informatics. https://doi.org/10.5445/IR/1000011477

[254] Walesiak, M. (2008). Cluster analysis with clusterSim computer program and R environment. *Acta Universitatis Lodziensis. Folia Oeconomica*, 216, 303–311. https://hdl.handle.net/11089/16186

[255] Walesiak, M., & Dudek, A. (2020). The choice of variable normalization method in cluster analysis. *Proceedings of the 35th International Business Information Management Association Conference (IBIMA)*, 325–340. https://bit.ly/3bjheVo

[256] Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open, 3*(8), Article e002847. https://doi.org/10.1136/bmjopen-2013-002847

[257] Wang, W., Guedj, M., Bertrand, V., Foucquier, J., Jouve, E., Commenges, D., et al. (2017). A Rasch analysis of the Charcot-Marie-Tooth Neuropathy Score (CMTNS) in a cohort of Charcot-Marie-Tooth type 1A patients. *PLOS ONE*, *12*(1), e0169878. https://doi:10.1371/journal.pone.0169878

[258] Watson, R., van der Ark, L. A., Lin, L.-C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in clinical practice: Item response theory. *Journal of Clinical Nursing*, *21*(19pt20), 2736–2746. https://doi.org/10.1111/j.1365-2702.2011.03893.x

[259] Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

[260] Wickham, H. (2017). *tidyverse: Easily install and load the 'Tidyverse'* (R Package Version 1.2.1) [Computer software]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=tidyverse

[261] Wind, S & Hua, C. (2021). *Rasch measurement theory analysis in R: Illustrations and practical guidance for researchers and practitioners*. Blockdown. https://bookdown.org/chua/new_rasch_demo2/
[262] Wit, E., van den Heuvel, E., & Romeijn, J.-W. (2012). 'All models are wrong...': An introduction to model uncertainty. *Statistica Neerlandica, 66*(3), 217–236. https://doi.org/10.1111/j.1467-9574.2012.00530.x
[263] Woo, S. E., Jebb, A. T., Tay, L., & Parrigon, S. (2018). Putting the "person" in the center: Review and synthesis of person-centered approaches and methods in organizational science. *Organizational Research Methods, 21*(4), 814–845. https://doi.org/10.1177/1094428117752467
[264] Wu, R., Wu, H., & Wang, C. L. (2021). Why is a picture 'worth a thousand words'? Pictures as information in perceived helpfulness of online reviews. *International Journal of Consumer Studies, 45*(3), 364–378. https://doi.org/10.1111/ijcs.12627
[265] Yalçın, S. (2018). Multilevel classification of PISA 2015 research participant countries' literacy and these classes' relationship with information and communication technologies. *International Journal of Progressive Education, 14*(1), 165–176. https://doi.org/10.29329/ijpe.2018.129.12
[266] Yan, W., Zhang, Y., Hu, T., & Kudva, S. (2021). How does scholarly use of academic social networking sites differ by academic discipline? A case study using ResearchGate. *Information Processing & Management, 58*(1), Article 102430. https://doi.org/10.1016/j.ipm.2020.102430
[267] Yıldırım, S. (2012). Teacher support, motivation, learning strategy use, and achievement: A multilevel mediation model. *The Journal of Experimental Education, 80*(2), 150–172. https://doi.org/10.1080/00220973.2011.596855
[268] Zhang, P., Aikman, S. N., & Sun, H. (2008). Two types of attitudes in ICT acceptance and use. *International Journal of Human-Computer Interaction, 24*(7), 628–648. https://doi.org/10.1080/10447310802335482
[269] Zhang, Z., Abarda, A., Contractor, A. A., Wang, J., & Dayton, C. M. (2018). Exploring heterogeneity in clinical trials with latent class analysis. *Annals of Translational Medicine, 6*(7), 119–119. https://doi.org/10.21037/atm.2018.01.24
[270] Zhang, Y., Sheng, M., Zhou, R., Wang, Y., Han, G., Zhang, H., Xing, C., & Dong, J. (2020). HKGB: An inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. *Information Processing & Management, 57*(6), Article 102324. https://doi.org/10.1016/j.ipm.2020.102324
[271] Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change, 21*(2), 245–266. https://doi.org/10.1007/s10833-019-09367-x
[272] Zou, X. (2020). A survey on application of knowledge graph. *Journal of Physics: Conference Series, 1487*, Article 012016. https://doi.org/10.1088/1742-6596/1487/1/012016
[273] Zuur A. F., Ieno E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*, 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

List of Publications

Lezhnina, O., & Kismihók, G. (2020). A multi-method psychometric assessment of the Affinity for Technology Interaction (ATI) Scale. *Computers in Human Behavior Reports, 1*, Article 100004. https://doi.org/10.1016/j.chbr.2020.100004

Lezhnina, O., & Kismihók, G. (2022a). Combining statistical and machine learning methods to explore German students' attitudes towards ICT in PISA. *International Journal of Research & Method in Education, 45*(2), 180–199. https://doi.org/10.1080/1743727X.2021.1963226

Lezhnina, O., & Kismihók, G. (2022b). Latent class cluster analysis: Selecting the number of clusters. *MethodsX, 9*, Article 101747. https://doi.org/10.1016/j.mex.2022.101747

Lezhnina, O., Kismihók, G., Prinz, M., Stocker, M., & Auer, S. (2022). A scholarly knowledge graph-powered dashboard: Implementation and user evaluation. *Frontiers in Research Metrics and Analytics, 7*, Article 934930. https://doi.org/10.3389/frma.2022.934930

List of Abbreviations

AIC	Akaike's Information Criterion
AIPS	Automated Item Selection Procedure
API	Application Programming Interface
ARI	Adjusted Rand Index
ASW	Average Silhouette Width
ATI	Affinity for Technology Interaction (scale)
AUC	Area under the Curve
BIC	Bayesian Information Criterion
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CSS	Cascading Style Sheets
CTT	Classical Test Theory
EFA	Exploratory Factor Analysis
ESCS	Economic, Social, and Cultural Status
FA	Factor Analysis
FAIR	Findable, Accessible, Interoperable, Reusable
HLM	Hierarchical Linear Modeling
HTML	HyperText Markup Language
ICC[4]	Intraclass Correlation Coefficient
ICILS	International Computer and Information Literacy Study
ICL	Integrated Completed Likelihood

[4] Item Characteristic Curve is also commonly abbreviated as ICC; for the clarity, only Intraclass Correlation Coefficient is abbreviated as ICC in this work, and Item Characteristic Curve is not.

ICLUST	Item Clustering
ICT	Information and Communication Technology
IRT	Item Response Theory
JS	JavaScript
KG	Knowledge Graph
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy
LC(C)A[5]	Latent Class (Cluster) Analysis
MAP	Velicer's Minimum Average Partial
ML	Machine Learning
MLR	Maximum Likelihood with Robust standard errors
MSA	Mokken Skaling Analysis
OECD	Organisation for Economic Cooperation and Development
ORKG	Open Research Knowledge Graph
PAF	Principal Axis Factoring
PCA	Principal Component Analysis
PISA	Programme for International Student Assessment
RF	Random Forest
RMSEA	Root Mean Square Error of Approximation
ROC	Receiver Operating Characteristic curve
RQ	Research Question
RSM	Rating Scale Model
SKG	Scholarly Knowledge Graph
SRMR	Standardized Root Mean Square Residual
UI	User Interface

[5] As I explained in section 2.3.2, LCCA is called LCA when discussed from the statistical perspective rather than from the perspective of clustering. To avoid confusion, in this work I call the method LCCA.

UEQ-S	User Experience Questionnaire (Short version)
VSS	Very Simple Structure

List of Tables

List of Figures