MASTER THESIS

Navigating Exams: Identifying Test-Taking Navigation Behaviour

DEPARTMENT OF INFORMATION AND COMPUTING SCIENCES
FACULTY OF SCIENCE

10/06/2021

Author
W.T.H. van Bakel

First Supervisor
Dr. M.J.S. Brinkhuis

Second Supervisor
Dr. Ir. J.M.E.M. van der Werf

First External Supervisor
J. Koops, MSc

Second External Supervisor
E. de Schipper, MSc

Utrecht University

I. Abstract

The digitalization of the educational domain has led to the availability of more data on students. Instead of only focussing on student performance, researchers and teachers nowadays are able to analyse student behaviour during educational tasks. Many studies explore this behaviour with the concept of item response time. Research on student navigation behaviour is lacking. This research extends the exploration of educational data and addresses specifically the analysis of navigational test behaviour.

With the help of the ACET 2018 high-stakes test dataset and the literature study, this research proposed measures for navigational test activities, including *steps, jumps, skips, changes* and *hops*. Additionally, we made a distinction between the navigation activity before and after reaching the last item. We used k-means clustering on a total of 21.565 students and their tests and identified three consistent test-taking navigation strategies. It was found that the majority of students mostly follow the linear order of the tests, the second group used more activities, and the last group was more active after the last item. No differences in the mean of student ability were found using the clusters.

Our results suggested there is little relation between the navigational activities and student ability, and difficult items were *skipped* and *changed* more often. Predetermined 'easy' items with a higher number of non-linear activities could serve as an indicator for hard-to-understand questions. As we discovered that most answer changes were advantageous, this research advocates for the use of free navigation in tests, giving the students an extra chance to show their full potential.

II. Acknowledgements

I have been blessed with the support I got from my supervisors, colleagues, family, girlfriend, and friends. Without their support, I would have never been able to finish this research, especially during this unusual year.

First, I want to thank my advisor, Matthieu Brinkhuis. As cheesy as it might sound, it is for Matthieu and his passion for teaching and the domain that I became enthusiastic in data analytics and data science. I even decided to follow the extra Applied Data Science profile and tried to follow as much data analytics as I could. For this research, Matthieu set me up with this research internship at Cito, and both supported and enthused me continuously throughout my research. Secondly, although I did not speak to my second supervisor, Jan Martijn van der Werf, that much during this research, I would like to thank him as well. Like Matthieu, Jan Martijn is a teacher that the University of Utrecht should be proud of. His passion for the development in IT, but even more for the students, is remarkable. I met Jan Martijn in my bachelor, and he always greeted me by name and made time to check up on me.

Next, I would like to thank my daily supervisors or colleagues at Cito, especially Jesse Koops and Eva de Schipper. I am incredibly grateful for both their academic and mental support during this research. In our pleasant weekly calls, they helped me shape this research, but above all, they were always first interested in how I was doing. In times of social isolation, it was nice to talk to people that said it was okay to take a step back. I wish them both good luck in their career and academic adventures.

Finally, I must express my gratitude to my family, girlfriend and friends who supported me throughout my years of study and through writing this thesis in an unusual year.

Thank you.

Wessel van Bakel

III. Table of Contents

I. Abstract 2
II. Acknowledgements 3
III. Table of Contents 4
IV. List of figures, tables and appendices 5
V. List of abbreviations 6

Chapter 1 Introduction 7
1.1. Context 8
1.2. Research Approach 8
1.3. Thesis Outline 11

Chapter 2 Literature Study 12
2.1. Background 12
2.2. Navigation Behaviour 13
2.3. Navigation Data 25

Chapter 3 Methods 29
3.1. Data understanding 29
3.2. Data Preparation 30
3.3. Method 32

Chapter 4 Results 33
4.1. Overall 33
4.2. Group 34
4.3. Student 37
4.4. Items 40

Chapter 5 Conclusion 42

Chapter 6 Discussion 45
6.1. Limitations 45
6.2. Future Research 46

7. References 47
8. Appendices 54

IV. List of figures, tables and appendices

Figures

Figure 1 Adaptation of Chapman's CRISP-DM Methodology 9
Figure 2 Depiction of the Research Sub Questions. 11
Figure 3 Number of papers about Educational Data Mining/Learning Analytics terms in Google Scholar by year (Romero & Ventura, 2020) 13
Figure 4 Navigational patterns suggested for measurements by Canter et al. (1985) 16
Figure 5 Flimsy (First left) and Laborious (Third left) navigation styles (Herder & Juvina, 2004) 16
Figure 6 Educational navigation patterns suggested for measurements (Bousbia et al., 2010) 19
Figure 7 Gall & Hannafin's framework for the study of Educational Hypertext System 19
Figure 8 Interface of LMS (left) and test-taking system (right) 21
Figure 9 Relationship between the amount of data and the level of granularity (Romero & Ventura, 2020) 26
Figure 10 Adapted relationship between the amount of data and the level of granularity for navigational test-taking 26
Figure 11 The Multistage Testing Structure in ACET 2018 30
Figure 12 The Structure of the ACET 2018 dataset 30
Figure 13 Time spent on test grouped by students only viewing item once or more 33
Figure 14 The distributions and averages of the navigation activity occurrences. 34
Figure 15 Groups of students clustered using before-last-item navigation measures 35
Figure 16 Groups of students clustered using after-last-item navigation measures 35
Figure 17 Groups of students clustered using all navigation measures 36
Figure 18 The statistics on student performance by clusters on all navigation measures 38
Figure 19 Boxplot of Navigation Activities per Items 41

Tables

Table 1 Research Methods related to Research Sub Questions grouped by CRISP-DM steps .. 11
Table 2 Measures for User Navigation presented by Herder & Juvina (2004) 17
Table 3 Types of Answer Changes Made to Items of different difficulties (Jacobs, 1972) 23
Table 4 Test-taking navigation activities distinguished in the data 31
Table 5 Overview of the total descriptive statistics on student's navigation behaviour 33
Table 6 Correlation between navigation measures and student ability 37
Table 7 Correlation between navigation measures and item score 39
Table 8 Types of Answer Changes Made to Items of different difficulties in our dataset 39
Table 9 Correlation between navigation measures and item difficulty 40
Table 10 Correlation between navigation measures and item presentation position 40

Appendices

Appendix A Detailed version of this research adaptation of Chapman's CRISP-DM 54
Appendix B Overview EDM application categories 55
Appendix C Description of variables found in cleaned data 56
Appendix D Visualization of student's test-taking navigation behaviour 57
Appendix E Snippet of our k-means algorithm code in Python 58
Appendix F Groups of students clustered using all navigation measures, divided into 4 sections based on item position 59
Appendix G Count per groups of clustered tests per student 60
Appendix H Types of Answer Changes Made to Items of different difficulties in our dataset per clusters 61

V. List of abbreviations

ACET	*Adaptieve Centrale Eindtoets* (Adaptive central final test)
CAT	Computerized Adaptive Testing
CRISP-DM	Cross-Industry Standard Process for Data Mining
CvTE	*College voor Toetsen en Examens* (The board of Tests and Examinations)
EDM	Educational Data Mining
EHS	Educational Hypermedia System
HLS	Hypermedia Learning System
IEA	International Association for the Evaluation of Educational Achievement
ITS	Intelligent Tutoring System
PISA	Programme for International Students Assessment
LAK	Learning Analytics & Knowledge
LMS	Learning Management System
MOOC	Massive Open Online Course
OECD	Organisation for Economic Co-operation and Development
RQ	Research Question
SQ	Sub Research Question
SSQ	Sub Research Question
UIDE	User Interface Design Environment

Chapter 1

Introduction

In recent years, a seismic shift has been happening within the secondary school education domain (Paraskeva et al., 2008). Secondary schools are increasingly integrating technology within their education to support the learning process. Additionally, with the advent of the pandemic, schools are forced to take an even faster digital leap in their everyday practices (Iivari et al., 2020). Classrooms are relocated to digital learning environments, and pencil and paper tests are replaced by laptops and digital tests. When in-classroom education is suspended for an extended period, digital teaching is no longer an option but a necessity (Dhawan, 2020).

This transformation in education leads to the creation of a vast trove of data. In the right hands, this data could prove a basis for better insight into student ability, cognitive style, adaptive feedback for teachers, tailored learning and improved tests. Institutions such as the International Association for the Evaluation of Educational Achievement[1] (IEA) and the Organisation for Economic Co-operation and Development[2] (OECD) make their assessment data publicly available to accelerate the rate of new discoveries. For example, OECD's international 2000 PISA survey was part of a policy change in Germany and helped to reach an above-average performance in reading, mathematics and science proficiency (OECD, 2013).

On a lower level, rich assessment data is recorded by most current digital learning environments. However, most of the data remain unused due to a lack of accessibility (Marsh et al., 2006). Historically, the education domain has been fixated on the assessment outcome (Stadler et al., 2019), but recent developments have researchers examine process data stored in computer-generated event logs. These logs are records of actions taken while working on a computerized assessment (Bunderson et al., 1988). Many researchers explore the concept of item response time (van der Linden et al., 2007; Vida et al., 2021; Yamamoto & Everson, 1997), although the meaning is not entirely clear due to the multiple possible causes for response time. A fast response could, for example, be either an effect of a skilled or unmotivated student (von Davier, Mullis, & Martin, 2020).

However, process data is more than just item response time, and much more fine-grained data are captured in digital education environments (Bezirhan, von Davier, & Grabovsky, 2020). Several studies dive deeper into educational process data and are measuring other attributes such as item strategy use (Greiff et al., 2016;

[1] https://www.iea.nl/

[2] https://www.oecd.org/

Salles et al., 2020; Shu et al., 2015), item revisits (Bezirhan et al., 2020) and navigational behaviour (Graf et al., 2010). Educational navigation behaviour is concerned with the movement of students interacting with an educational system and utilizes temporal sequence data. In many learning systems, analysis of navigational behaviour could mean the sequence of the visited course pages. Which pages are (most) visited or which pages are revisited? The navigation behaviour of students in learning systems has been actively studied in recent years (Pechenizkiy et al., 2009), but research specifically in digital test-taking systems is lagging behind.

This research seeks to find patterns in student's navigation behaviour in high-stakes digital tests to address this vacuum. Therefore, the main objective of this research project is to continue examining the use of educational process data and explore, in particular navigational test-taking behaviour, using a dataset of high-stakes tests.

This study could provide new knowledge for educational institutions as well as for the field of psychometrics by:

- Discovering differences in navigation behaviour and finding navigation strategies
- Discovering the relation to student ability
- Supporting the research for student cognitive style
- Outline difficult or poorly written/structured items
- Change what data is stored from digital assessment and how it is stored

1.1. Context

This research tries to understand student navigation behaviour in test-taking environments in collaboration with Cito[3], a Dutch agency that supports governments, institutions and awarding bodies developing testing and monitoring systems (Cito, 2021). Cito provided exam-related data produced by the computer-based testing system, Facet[4]. Facet is a governmental platform responsible for central digital examinations and tests in secondary education. Facet utilizes the test and test items developed by Cito and sends test data back for test improvement and learning analytics.

1.2. Research Approach

Due to the absence of specific navigation behaviour theory in the educational domain, we will follow an inductive research approach, also known as a bottom-up or data-driven approach. With the help of the cross-industry standard process for data mining (CRISP-DM), this research tries to answer research questions by

[3] https://www.cito.com/

[4] https://www.duo.nl/zakelijk/voortgezet-onderwijs/examens-en-diplomas/facet/

analysing observations in the dataset. The CRISP-DM methodology is a successful and practical data mining standard, with six organized and clearly defined phases, allowing easy understanding and revision of a project (Chapman et al., 2000). This research uses an adaptation of the CRISP-DM, depicted in Figure 1 Adaptation of Chapman's CRISP-DM and more detailed in Appendix A, as this research does not share the business narrative with CRISP-DM, and there currently will not be a deployment of any solution.

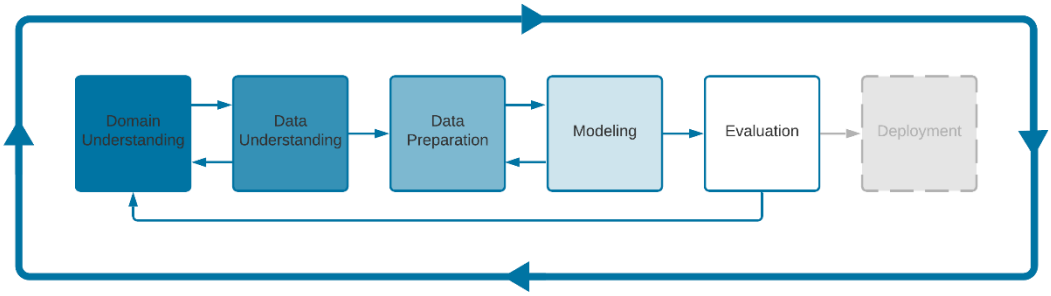


Figure 1 *Adaptation of Chapman's CRISP-DM Methodology*

The initial phase is changed from Business Understanding to Domain Understanding. Instead of solely looking for project goals, this phase is extended with a literature study in the Educational Data Mining and Learning Analytics domain to discover the state-of-the-art and search for educational (test-taking) navigation behaviour. The following two phases revolve around the dataset. After finding the attributes that contribute to navigation behaviour, we transform the log data into a better analysable pattern format. Next, this research will devise navigation measures and use a data mining technique that will identify different navigation behaviour styles. Lastly, we evaluate the models' results of the state-of-the-art techniques and evaluate our model by testing it on a new similar dataset.

1.2.1. Research Questions

In alignment with the problem statement and the obtained dataset, this research defines the following main research question (RQ).

RQ *What can we learn from digital test-taking navigation behaviour?*

This research will attempt to answer this main question through several sub-questions (SQ) and their respective answers. Figure 2 depicts the total research objective.

SQ1 *What different test-taking navigational behaviour exist?*

- ***1.1*** *What different digital navigational behaviour exists outside the educational domain?*
- ***1.2*** *What different digital educational navigational behaviour exist?*
- ***1.3*** *What different digital navigational behaviour exist in test-taking?*

To better understand what we can expect from the data, we aim to discover literature in navigational behaviour. We will start with broad

research outside the educational domain and funnel down subsequently to research in the educational domain and end with specific test-taking navigation behaviour research.

SQ2 *What techniques can be used for inferring navigational behaviour from digital tests?*

- ***2.1*** *What data can be used for interpreting navigation behaviour?*
- ***2.2*** *How can data mining techniques be used for analysing digital navigation behaviour?*

By virtue of the bottom-up approach, we attempt to get new insights by analysing the data and compare this later with the current theory. However, to understand what data we should use and how we can infer navigational behaviour from the data, we examine the analysis of closely related studies. Accordingly, the data will be prepared and examined to our needs.

SQ3 *What type of navigational behaviour can we discover in the provided data set?*

- ***3.1*** *Which measures can be used to analyse digital navigation behaviour?*
- ***3.2*** *Which different navigation strategies can be identified based on navigational behaviour?*

With this question, we will analyse the navigation behaviour and try to identify various student navigational characteristics. In addition, we will try to discover the differences in navigation behaviour and the possibility to group students with the same behaviour.

SQ4 *How does navigational test behaviour relate to student performance?*

- ***4.1*** *How does navigational test behaviour relate to student ability?*
- ***4.2*** *How does navigational test behaviour relate to answering a question correctly?*

In this section, we attempt to identify the relationship between navigational behaviour and the student's level of ability. Additionally, we analyse the existence of an optimal navigation strategy. Finally, in sub-question 4.2, we will analyse how navigational behaviour will impact the score on a single item.

SQ5 *How does navigational test behaviour relate to a test and its items?*

To conclude, this research analyses the relation between navigational test behaviour and items. We will answer questions like 'Do item difficulty or item presentation order have any correlations with navigation behaviour?' to improve the test and its respected items.

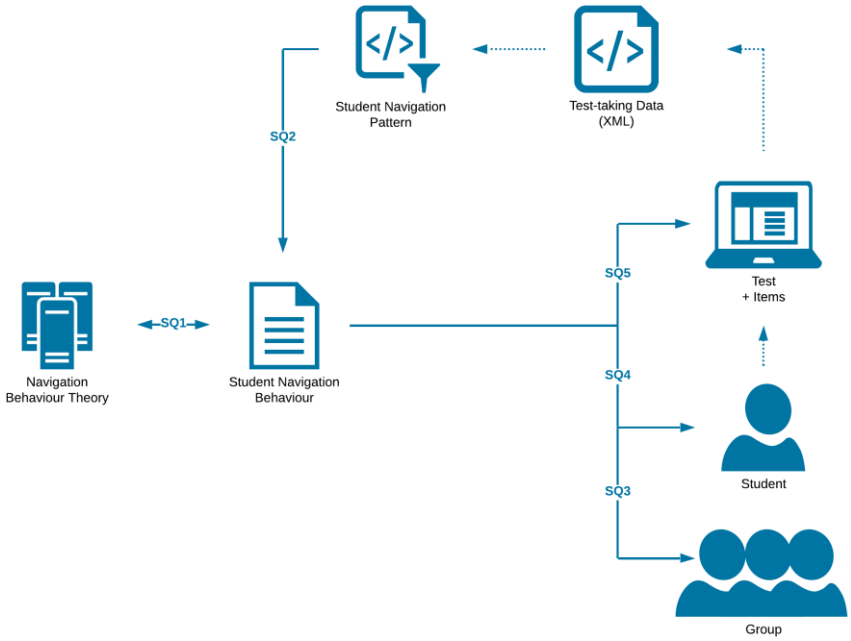


Figure 2 *Depiction of the Research Sub Questions.*

1.2.2. Research Method

We used different research methods to answer the research questions. Table 1 shows the research methods used in this study to answer the research questions. To better understand the domain and discover opportunities in other research, we begin with a literature study to answer the first two sub-questions. Finally, we answer sub-question 3, 4 and 5 by analysing and modelling the data.

Table 1

Research Methods related to Research Sub Questions grouped by CRISP-DM steps

	Domain Understanding		Modelling / Evaluation		
Research Method	SQ1	SQ2	SQ3	SQ4	SQ5
Literature Study	✓	✓			
Data Analysis			✓	✓	✓

1.3. Thesis Outline

The rest of this paper is organized as follows. In Chapter 2, we identified the gap in the state-of-the-art education data mining/ learning analytics domain. Then, in chapter 3, we funnelled down to test-taking navigation behaviour in our literature study. Chapter 4 describes the method and measures used in our research, and in chapter 5, the results are discussed. Finally, in chapter 6 and 7, we draw a conclusion and discuss limitations and potential future research.

Chapter 2

Literature Study

Following the first step of the CRISP-DM, we got a better understanding of our research domain by reviewing systematic literature reviews by other researchers. Subsequently, we used our research questions to guide us in first understanding navigation behaviour and secondly discovering the state-of-the-art educational data mining.

2.1. Background

The interest in educational data and its analysis are growing at a fast pace. Where in 2008, there was not a single publication tagged “learning analytics” or ‘educational big data’, years later, in 2017, the number of publications had increased by more than 2.000 (Hwang et al., 2018). Several research communities have grown due to the collective interest in digital educational data, of which the two most prominent ones are Educational Data mining[5] (EDM) and Learning Analytics & Knowledge[6] (LAK). The numbers of papers published by these communities have similarly grown exponentially (Figure 3). The EDM community is concerned with developing methods for exploring the unique types of data that come from educational environments (Bakhshinategh et al., 2018). During the first LAK conference, learning analytics was defined as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs” (*LAK ’11: Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, 2011). Both communities share the goal of enhancing educational practice (Calvet Liñán & Juan Pérez, 2015) but have slight differences and different viewpoints (Romero & Ventura, 2020). That is why this research will use these communities, their respective conferences, journals as the start of our literature study for exploring educational navigation behaviour.

In addition, Hwang makes a case for exploring and employing learning analytics in various application domains, particularly those seldom investigated ones (Hwang et al., 2018). Therefore, this research explores educational navigation behaviour in test-taking environments. At first glance, we see a small though gaining interest in navigation behaviour and process mining, with navigation studies in E-book environments (Ogata et al., 2017) and Learning Management System (Aldowah et al., 2019). However, test-taking navigation behaviour looks to be untouched. This

[5] https://educationaldatamining.org/
[6] https://www.solaresearch.org/events/lak/

could affirm that our research tries to explore new opportunities within the EDM and LAK domain; however, it will have little guidance in doing so.

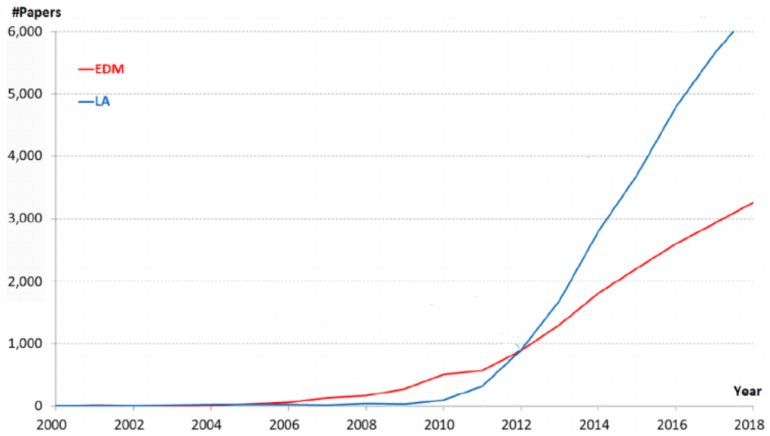


Figure 3 *Number of papers about Educational Data Mining/Learning Analytics terms in Google Scholar by year (Romero & Ventura, 2020)*

2.2. Navigation Behaviour

To identify test taker's navigation behaviour, we intend to discover behavioural indicators by incorporating domain knowledge and analysing patterns in the event logs. Because behaviour within a test has not been the most researched field, this research tries to learn from studies outside the educational test-taking scope. In the following section, we will start by describing interesting general navigation behaviour and measures. Next, we will narrow down our search, going layers deeper into digital, educational and, finally, test-taking navigation behaviour.

"Behaviour is the way that a person, an animal, a substance or other behave in a particular situation or under particular conditions" (Cambridge University, n.d.). If this behaviour revolves around following routes or paths, navigational behaviour is at play. Navigational behaviour exists in many different forms, and by entering the term in Google Scholar, over 18.000 thousand different studies pop up. Thuring et al. (1995) considered that navigation has two aspects: direction (forward/backwards) and distance (step/jump). Most of the early navigational studies primarily investigate the spatial/habitual movement behaviour of animals and humans (Bingman et al., 1990; Carr, 1965; Streeter et al., 1985), but with the rise of modern technology, a new domain has been added. The combination of the significant presence of humans online, the endless possibilities in the digital world and the recording of human-computer interactions have led to the interest in digital navigation behaviour. To optimize user experience (F.-H. Wang & Shao, 2004) and maximize return (Baumgarten et al., 2000), researchers and companies try to learn as much as possible about how and why users navigate/browse through (commercial) hypermedia systems, like web pages.

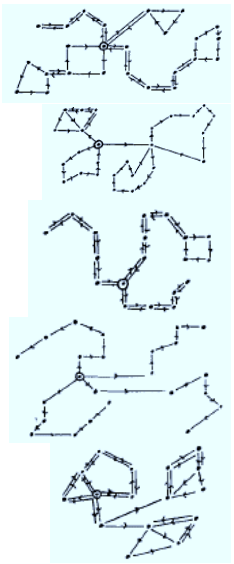
2.2.1. Digital Navigation Behaviour

There is a vast amount of literature on web navigation and individual navigation differences. For example, Kralisch et al. (2005) discovered that users' cultural backgrounds influence their website navigation patterns. However, the factors that have the most impact on navigation behaviour are the (i) user's goal, (ii) user's expertise, (iii) user's cognitive style, (iv) structure of the system and (v) technological aspects (Herder & van Dijk, 2004). A user could navigate with a goal-directed task, explore available information with more unstructured browsing or perhaps have a combination of different goals during a single session. Differences by expertise were discovered by Eveland & Dunwoody (1998), where content/system novices tend to navigate linearly when it is made available. On the other hand, experts tend to make use of a non-linear structure in hypermedia systems. Graff (2005) focused on the difference in navigation pattern by looking at cognitive style. With relatively simple measures like number and proportion of (re)visited pages, he discovered differences in web browsing strategies between individuals with verbalisers and imager cognitive styles – that is, the preference of learning through text or image.

The most cited problem within digital navigation is disorientation (Thuring et al., 1995). Disorientation happens when users do not know where they are, how they came there and where they should go next. Following this problem, several researchers have identified three general navigational profiles of internet users (Barab et al., 1997; Lawless et al., 2002; Niederhauser et al., 2000). Knowledge seekers, feature explorers and lastly, apathetic hypertext users, who neither care about gathering information nor explore system's features. Catledge & Pitkow (1995) took another take on navigation pattern and classified users into different browsing strategy groups by analysing the frequency and depth in their navigation pattern. The relationship between frequency of visited pages and length of navigation path resulted in the labelling of specific navigation patterns as random "serendipitous browsing", "General Purpose Browsing", and lastly, the goal-oriented "search browsing".

Research on web navigation often refers to the popular typology defined by (Canter et al., 1985). The authors of one of the most studied studies in this domain proposed a taxonomy that suggested five navigation strategies:

- *Scanning* covering a large area, however without paying attention to detail
- *Browsing* navigating until a user's interest is caught
- *Searching* seeking something (page/file/document) with vast motivation
- *Exploring* Understanding the extent and nature of the field
- *Wandering* navigating in an unstructured way without a strategy



With the help of Alty's discussion of path algebras and connected graph (1984), Canter et al. were one of the first to model and analyse the user's behaviour based on their actions during a digital session. Each state of the system became a "node", and "paths" were transitions to other states. Inspired by humankind's daily movement ("journeys"), this research investigated patterns in user's digital path between nodes.

> *"Consider the navigation of the surroundings involved in a person's movements during a day. Several locations will be visited. The person will typically start and end at the same place (his home) but, during his travels, he will make smaller excursions from and back to other places. Often these excursions will be nested (a trip from home to the office, to the bar of the local pub at lunchtime, from the bar to the toilet, then the telephone, then back to the bar, back to the office, to a newsagent on the way home, and finally back to his home."*
>
> ***(Canter et al., 1985)***

A total of six indices were used to characterise the graph associated with the navigation path. Unfortunately, most measures are defined abstractly and measured in an unknown manner. We were unable to understand how they realised the scores, and therefore this research is unable to apply Canter's measures in our work. However, it is still interesting to examine these measures and how they contribute to the final identification of navigation strategies. The first four indices revolve around the score of a pattern found in the user's navigation behaviour (Figure 4). A high score for a pattern indicates it is found frequently within that user's navigation. The first pattern is a route that does not visit the same node twice, the *path.* Second, a *ring* is a route that returns to the starting node. A special kind of ring is the *loop* that does not cross any nodes, besides the start node, twice. Lastly, the *spike* is a route that does cross every node twice and

retraces the original path on its return. The last two indices are ratios concerning the numbers of nodes visited. The first one provides an indication of the range of exploration undertaken and is measured by dividing the number of nodes visited by the total number of nodes in the system. The other ratio is the number of different nodes visited to the total number of paths, which measures the redundancy of the strategy the user has utilized. With these six indices, Canter et al. (1985) characterised the five navigation strategies mentioned before.

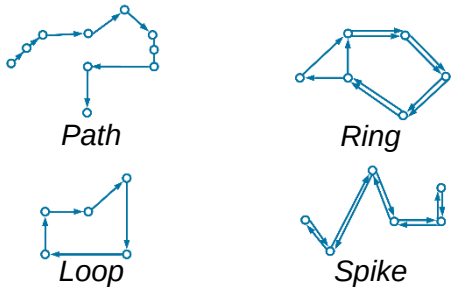


Figure 4 *Navigational patterns suggested for measurements by Canter et al. (1985)*

Herder and Juvina (2004), too, understood that user navigation paths could be modelled as graphs and aimed at finding patterns in user navigation that indicate a user's vulnerability to perceive disorientation. They were able to identify two navigation styles: "flimsy" and "laborious" navigation (Figure 5**Figure 5**). Flimsy navigation arose as a weak navigation style and was best characterized by relatively short navigation paths and a low number of cycles. Users who utilize such a style tend to have low scores on an active mood, working memory and locus of control (Herder & Juvina, 2004). On the other hand, users with a laborious navigation style understand the website's structure fast and rarely experience disorientation.

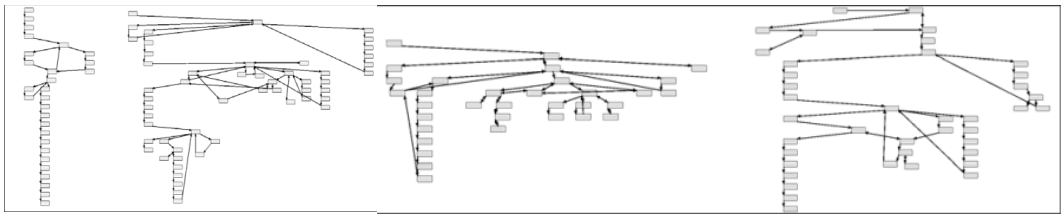


Figure 5 *Flimsy (First left) and Laborious (Third left) navigation styles (Herder & Juvina, 2004)*

We present their study since they utilized graph-theoretic and statistical methods for analysing their data and made use of a selection of navigation measures that are useful for describing patterns within a user's full navigation path. The three main measures are revisits, view time and complexity. Revisits are very common in web navigation; view times are reported to be an important indicator, and navigation complexity reflects the cyclical structure of the navigation graph and the length of navigation sequences within the graph. The full description of measures can be found in Table 2 on the next page.

Table 2

Measures for User Navigation presented by Herder & Juvina (2004)

Overall Measure	Measure	Description
Number of Pages and Revisits		
	Path length	The total number of requested pages by the user in a single session
	Relative amount of revisits	The probability that any page visited has already been viewed as suggested by Tauscher & Greenberg (1997)
	Page return rate	The average number of visits to all pages that have been visited at least twice.
	Back button usage	The percentage of back button clicks of all navigation actions
	Relative amount of home page visits	The percentage of home page visits based on all page visits
View Times		
	Average view time	The average view time of a page
	Median view time	The median view time of a page
Navigation Complexity		
	Fan degree	The ratio of the number of links followed and the number of unique pages visited as suggested by (Rauterberg, 1999)
	Number of cycles	The difference between the number of links followed and the number of pages visited as suggested by Rauterberg (1999)
	Path density	Compares the user's navigation graph to its fullest possible connected graph as suggested by Rauterberg (1999).
	Compactness	Compares the average distance between any two pages in the navigation graphs to a theoretical minimum and maximum, as suggested by Mceneaney (2001).
	Average connected distance	The average length of a path between any two visited pages as suggested by (Broder et al., 2000)

2.2.2. Educational Navigation Behaviour

Although traditional paper books and tests are likely to stay with us for some years, digital/hypermedia systems are assuming significant roles in the way we educate students (McEneaney, 2001). Numerous hypermedia systems have recently been developed specifically for the educational domain. The arrival of these Educational Hypermedia Systems (EHS) advanced the ability of researchers/teachers to observe and understand student behaviour. When previously a researcher had to physically observe the learning process of a student, most EHS's record every user

activity and have the ability to export a variety of data. Identifying and understanding a student's behaviour, style or preference could lead to better design of systems and help to guide an individual with their (adaptive) learning process.

Papanikolaou and Grigoriadou (2004) categorised the learner's observable behaviour within EHS into three groups: performance, temporal and navigational behaviour. Performance is historically the most analysed behaviour in the education domain. Performance indicators, such as test scores and student ability, have always been the leading indicators in this domain. Temporal behaviour is obtained by measuring the time spent in different types of educational resources. Lastly, navigational behaviour can be captured by identifying the number of hits on educational resources and the sequence of visits.

In most EHS, the user, or in this domain, the learner, is given some control over their navigational behaviour, therefore allowing a wide variety of navigational possibilities and behaviour. (Bousbia et al., 2010), who researched educational navigation behaviour in Learning Management Systems (LMS), found similar navigational behaviour to the web navigation research of Canter et al. (1985). However, to fit these navigational characteristics to the educational domain, they proposed an adapted version of the navigation typology with four navigation types:

- *Overviewing*
  - covering a large portion of course pages
  - little time spend on individual pages
  - (close to Canter's *scanning*)
- *Flitting*
  - covering a large portion of course pages
  - little time spend on individual pages
  - with a lack of focus
  - (close to Canter's *wandering*)
- *Studying*
  - covering all or large portion of course pages
  - more time spent on individual pages
- *Deepening*
  - covering all or large portion of course pages
  - relatively long time spent on total course and course pages
  - using the web to obtain more information about course

Detecting these navigation types with the Canter's patterns (Figure 4) was discovered to be complicated. The learner usually goes to-and-from pages according to the learning task mixing different patterns (Bousbia et al., 2010). Their research added indicators for path linearity and the detection of central nodes to the pattern analysis and distinguished three educational browsing patterns (Figure 6). Path linearity is the ratio of the number of different pages divided by the number of steps. If this indicator is close to 1 in the interval [0,1], the navigation path is linear and identified as the pattern *scholar,* which is similar to Canter's *path*. For non-linear patterns, by analysing sub-sequences and identifying central nodes,

they distinguished the *star* pattern with several *loops*. If a pattern does not fall into any of these two groups, the pattern becomes the last pattern *dispersion*, indicating a learner moves in an unorderly fashion.

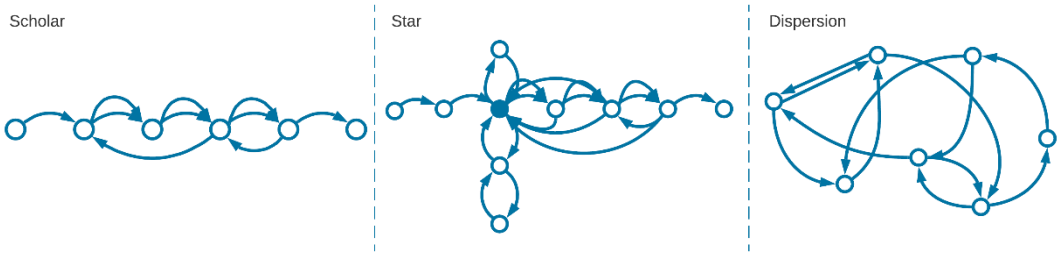


Figure 6 *Educational navigation patterns suggested for measurements (Bousbia et al., 2010)*

However, comparing a learner's (navigational) behaviour is far from trivial as it is impacted by several factors. Gall & Hannafin (1994) indirectly identified these factors for educational (navigation) behaviour when they developed one of the first frameworks to support the study of educational hypertext systems. Inspired by the information-seeking framework by Marchionini & Shneiderman (1988), Gall & Hannafin adapted the framework for the educational context and analysed the relationships among hypermedia learning system components. Similar to the previously discussed research by Herder & van Dijk (2004), the authors describe the relationships between the user's goal, the user's attributes and the structure of the system in an educational context (Figure 7**Error! Reference source not found.**). Although Gall & Hannafin (1994) did research learner's navigation control, their research had a more technological base. Their research focussed on the design of the hypermedia learning systems (HLS) and the boundaries determined by designers. They discovered four control structures allowed by most HLS at that time. *Searching*, seeking information specifically. *Browsing,* looking through information without a goal. *Connecting*, creating electronic links between information. Lastly, *collecting*, assembling information outside of the system.

Instead of improving the design of HLS, this research is focused on the learner itself. However, to fully understand this behaviour, we use their framework to get a grasp on the other components and their impacts on navigational behaviour.

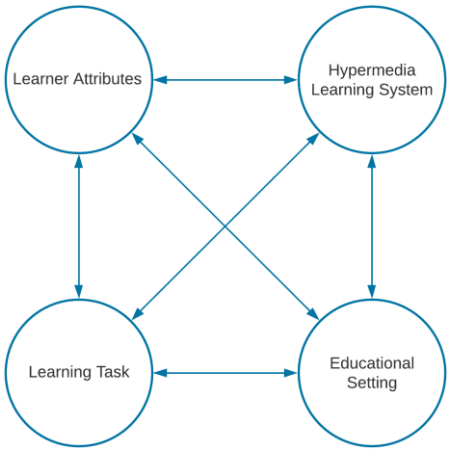


Figure 7 *Gall & Hannafin's framework for the study of Educational Hypertext System*

Different types of digital educational environments exist, and each one facilitates the use of different navigation patterns and produces different data. To contextualize the navigation behaviour, it is helpful to get familiar with some of the most prominent educational environments and understand the differences in these systems. Current research on educational navigation behaviour has been done mainly on LMS. The software can be utilized in several ways, but it mainly serves as a central location for one or multiple courses. It is a total package of software that provides course-delivery functions, including documentation, tracking, and reporting of training programs, classroom and online events, e-learning programs, and training content (Romero & Ventura, 2020). Modern examples of LMSs are Blackboard, Moodle and Canvas. LMSs are able to record any student activities on the platform, like reading, writing, commenting and taking tests.

Test-taking features also exist outside the context of an LMS in specialised educational testing systems, like the Facet System, used in this research. The main goal of such test systems is to measure the students' level of knowledge with respect to one or more concepts by using a series of questions/items (Romero & Ventura, 2020). Traditional test systems store students' answers, their final scores, but newer systems store all sorts of inputs, like mouse clicks and text input with timestamps.

The differences between these systems and their outputs can also be described using Gall & Hannafin's (1994) distinction of top-level features in virtually all digital systems: the knowledge base, the system’s interface and navigational facilities. The knowledge base is not only the collection of information in a specific context but also the structure and connections of the elements. In comparison to the LMS, elements in test-taking systems are relatively homogenous. Questions may be presented in different forms, but they remain a question element. Nevertheless, students could navigate in LMS to diverse elements such as documents, video tutorials, discussion posts and course pages. The connections between the elements in the knowledge bases differ for these systems as well. The hierarchical structure of elements in LMS provides the students with several ways to access information, including diving deeper into a subject (Gall & Hannafin, 1994). Questions in test-taking systems are often structured in a predetermined order, and connections between these elements are primarily linear.

Since the connections between elements in these systems differ, they also ask for different navigation tools. The navigation in test-taking mirrors the structure of its elements and provide the user with a linear layout, where students can move to the next or go back to previous questions. On top of that, most test-taking systems allow the users to get an overview page of all questions and jump immediately to the desired question. The LMS allows users to navigate more freely than test-taking systems. Users have the possibility to navigate linearly in some parts of the LMS (e.g. slides/documents), but the majority of e-learning environments make use of an interconnected layout (Makany et al., 2007). Whilst scrolling through pages and reading documents, the users have full navigational control by a combination of linear layout and an always-available central index page.

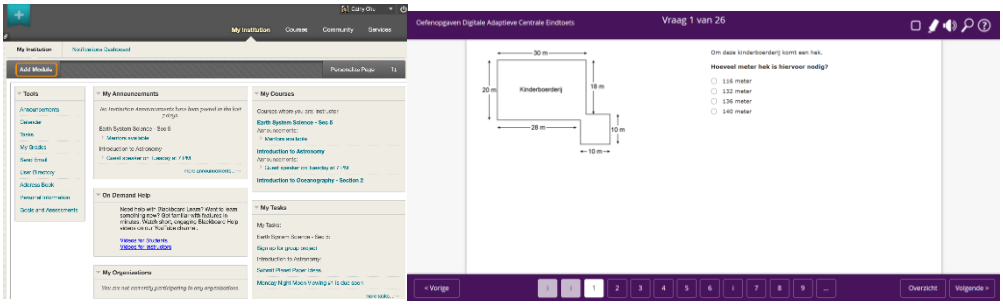


Figure 8 *Interface of LMS (left) and test-taking system (right)*

Differences in educational tasks must also be considered, as they provide context for the goal for a particular student. Educational tasks represent tasks “directly associated with the delivery of the educational component to students” (Mehdi Khosrow-Pour, 2010). They guide the students and activate and control the learning and assessment processes (Richter, 2012). Students may need to read through different elements to understand a concept, find specific/important information, collaborate with colleagues or take an assessment test. These tasks go hand in hand with the digital educational environment as they could serve as the function for how modern educational environments and their features are structured. For example, in contrast to goal-free learning, procedural learning asks for restriction of a student’s navigation control and a predefinition of concepts and steps (Gall & Hannafin, 1994). Strategies for learning tasks are widely researched and discovered to be different for learning tasks (Broekkamp & van Hout-Wolters, 2007). On the contrary, strategies for assessment tasks are lacking.

Besides the various educational systems and tasks, differences between students play a role regarding their navigation behaviour as well. These differences in attributes, like prior knowledge, cognitive attributes and learning styles, may affect the way in which a student interacts with learning systems.

Prior knowledge has been recognised as an important attribute. Macgregor (1999) discovered more purposeful navigation and better time allocation for students with greater domain knowledge. Akçapınar & Altun (2010) also found differences between students with low and high prior knowledge. With the help of McEneaney's metrics (2001), compactness and stratum, their research detected that students with higher prior knowledge had “better structured” navigation patterns. Additionally, students with less domain knowledge did not benefit from menu choices as much as experts (Juvina & Oostendorp, 2006).

Cognitive style is defined by Riding & Rayner (2013) as “the individual’s preferred and habitual approach to organizing and representing information.” Many cognitive learning style models exist in the literature, like the models by Witkin et al. (1977), Felder & Silverman (1988) and Riding & Cheema (1991). Research has shown that these cognitive styles correlate with performance in a Web-based environment (Tsianos et al., 2008; K. H. Wang et al., 2006). Additionally, according to Antonietti & Giorgetti (1998), one way to measure cognitive styles is through the navigation behaviour analysis of users. This could indicate that we could say something meaningful about student performance with the help of student’s navigational

behaviour. Graf et al. (2010) mention that so far in most approaches, navigational behaviour has been neglected in the detection of cognitive style but emphasizes that it could make the automatic detection process of learning styles more accurate. Another finding regarding the cognitive styles of Witkin et al. (1977) was that field-independent students use an active and analytic learning approach. On the other hand, field-dependent students tend to adopt a global approach and prefer guidance from the system (Chen, 2010).

2.2.3. Test-taking Navigation Behaviour

Finally, with the knowledge obtained from web and educational navigation, this research dives deeper into test-taking navigation behaviour. Test-taking navigation behaviour occurs when both the system and educational task revolve around the assessment of students' proficiency. Research in this particular type of navigation behaviour is lacking; nonetheless, there are some pioneers that investigated this subject.

As we mentioned earlier, the apparent difference between web navigation and HLS navigation is the structure of systems and elements. Where elements in websites and HLSs are hierarchically connected, elements in most test-taking systems either are linearly connected and not in a prearranged order. The lower dimensionality of test-taking systems limits the navigation tools (as seen in Figure 8) and possibilities and therefore limits student's control. Some test variants reduce the student's navigation options even further, only allowing the student to answer and proceed to the next question. Strict order navigation is commonly used in tests where immediate feedback is provided after each question (Vasilyeva et al., 2010) or in Computerized Adaptive Testing (CAT). In CAT, the next question is based on the student's answer to the previous question(s) and their current ability level (Wainer et al., 2000). Vasilyeva et al. (2010) studied and compared these navigational structures and discovered that students benefit from flexible order navigation. Students who were able to return to/revise earlier questions during tests distributed their time and efforts better. Additionally, their results suggest that strict order navigation tests, like CAT, could compensate their navigation limit by providing additional information for each question, like item difficulty or recommended response time. A student taking this test will benefit from immediate feedback adaptation without the negative impact caused by the fixed order navigation.

To identify test-taking navigation behaviour and strategy, we must understand the possible navigation activities in non-strict tests. The most common allowed actions include starting/submitting the test and moving forwards or backwards to questions through the next or back buttons. Moreover, on item level, it could be interesting to observe student actions. However, even less research has been done into this matter. Wu (2017) piloted a test eye-tracking application to observe cognitive processes, like problem-solving, and Salles et al. (2020) researched answering strategy in combination with tool usage.

Returning to the between-question navigation behaviour, several studies discovered that most students tend to first answer the questions in a linear order and flexibly navigate mainly at the end (Lee & Haberman, 2016; Vasilyeva et al., 2010). This flexible navigation generally occurs by students 'jumping' forwards or backwards, moving out of sequence, skipping one or more questions. The interpretation of the jumping forward could either be scanning the questions, which supports time management (although not at the end like mentioned or on the other hand, it could indicate disengagement, disinterest or poor planning (Jeske et al., 2014). Although Jeske et al. (2014) did not research the direct relationship between jumping and disengagement, they did find a negative correlation between both backward and forward jumping and test performance. Rindler (1980) discovered that middle ability students skip items more than their more and lesser able peers. She found that higher ability students benefit from skipping while lower ability students are negatively impacted by this activity, especially on tests where items get more difficult over time. Additionally, her research concludes that the relation between skipping behaviour and test score is not affected by the student's ability to manage time. Rindler (1980) suggests that physiological attributes like energy, motivation and frustration could play a role. Furthermore, other studies (Kim & Goetz, 1993; McClain, 1983; Stenlund et al., 2017) confirms skipping as an ineffective strategy for low achievers and suggest test anxiety as one of the main explanations.

Additionally, students going back to questions and revising their answer is an interesting test-taking (navigation) behaviour. Test-takers often get the advice "Go with your gut." or "Your first hunch/guess is usually the right one" (Lynch & Smith, 1972). However, in 1972, Lynch & Smith already said "research has been available for over forty years indicating that reconsidering test items tend to raise scores.". Both Jarrett (1948) and Jacobs (1972) discovered that even the low scoring test-takers improved their scores more by reconsidering items. Additionally, Jacobs' research (1972) on 50 students analysed the impact of item difficulty and discovered low and moderate difficult items benefit the most of answer changes (Table 3). The results in these early studies were later confirmed by researchers analysing navigational test data (Ferguson et al., 2002; Fischer et al., 2005; Mcnulty et al., 2007); students who change answers are more likely to change them from incorrect to correct. However, Jeon et al. (2017) discovered that item difficulty was generally greater when wrong answers were revised to correct answers.

Table 3

Types of Answer Changes Made to Items of different difficulties (Jacobs, 1972)

(n=735 students)	Item difficulty			
	Low	*Moderate*	*High*	Total
Right to Wrong	5.6%	6.8%	7.8%	20.2%
Wrong to Right	18.2%	25.0%	12.7%	55.9%
Wrong to Wrong	2.7%	7.9%	13.3%	23.9%
Total	26.5%	39.7%	33.8%	100%

Other test-taking strategies besides navigation are mentioned and categorised in the studies of Hong et al. (2006), Bıçak (2013) and (Stenlund et al., 2017). Overall, the strategies are categorised into three categories, structural (time management and navigation sequence), cognitive (in item strategy like eliminating answers) and motivational (guessing).

Research on test behaviour thus far has been primarily done in the structural category on item response time, which refers to the time used by a student to answer a test question (van der Linden, 2009). This is not surprising as next to the students' responses, item response times are the most widely collected data within computer-based tests. Unfortunately, item response time often has more than one possible cause and might not be able to explain test-taking behaviour without context (Bezirhan et al., 2020). However, the analysis of item response time has been found helpful in some areas, for example, for optimizing test assembly in nonadaptive tests, item selection in CAT (van der Linden, 2008), detecting cheating (van der Linden & van Krimpen-Stoop, 2003) and detecting rapid-guessing behaviour (Schnipke & Scrams, 1997; Wise et al., 2009).

Rapid-guessing behaviour is not solely based on short item response times as competent students could have short item response times but correct answers (Lee & Jia, 2014). In low-stakes tests, Wise & Kong (2005) discovered that rapid guessing had a negative impact on motivation. The results of the studies of rapid-guessing behaviour in high-stakes tests had a navigational angle. By examining the response time distributions for fast responses, the research of Schnipke (1995) and Schnipke & Scrams (1997) suggests that rapid-guessing behaviour is a function of item position. Students in high-stakes tests typically start with trying to find the solution to every item. However, as time expires toward the end of the test, students might switch to rapid guessing. As item position was found to be an important factor and tests tend to have different test versions, the items that are affected will differ across students and are hard to analyse on item level (Schnipke & Scrams, 1997).

In addition to the analysis of item response time in relation to item position, researchers examined temporal test-taking behaviour for different student characteristics, including performance and ethnicity (Lee & Haberman, 2016). This study confirms that most students start slowly and increase their pace towards the end and found that the overall test pace of higher-performing examinees was relatively more stable. Their results also indicate that students from different countries have different time-management strategies. European students tend to read information items more carefully and allocate their time evenly between items. On the other hand, Lee & Haberman (2016) mention that Asian students are more coached for tests, have more variable response times, tend to skip/skim items more and complete tests more often. Their research concludes with a recommendation. "It may also be beneficial to exploit the test design to skip items when needed and return later. Especially in the case of early items, this strategy would prevent someone from expending too much time unwisely on particular items when there is still ample time to spare." (Lee & Haberman, 2016). This is in

line with the recommendation of the previously mentioned research by (Vasilyeva et al., 2010).

2.3. Navigation Data

Instead of exploring literature about test-taking navigation behaviour, this chapter focuses on the practice of test-taking navigation analysis. This chapter builds on the differences found between educational systems and examines how educational data is presented. We intend to understand how we can infer navigational information from raw data. Additionally, we try to find techniques in recent EDM and LA literature that are suitable for our research.

2.3.1. Educational Data

The increase in digital educational environments has created large repositories of data reflecting how students perform and behave. (Romero & Ventura, 2010). Standardised learning and the ability to log student interactions and responses in online courses provide a gold mine of raw educational data (Mostow & Beck, 2006). The variety of data is immense as data differs between educational systems, can be captured from multiple sources, in different formats and with different levels of granularity (Romero & Ventura, 2010). Granularity is defined as the level of detail and differs between the multiple levels of meaningful hierarchy. At the lowest level of granularity, one can find detailed student interaction as mouse clicks and keystrokes. The lower the level of granularity, the more data is captured. As seen in Figure 9, Romero & Ventura (2010) use four levels of hierarchy as a simple representation for the educational context. However, in theory, the events/action level can be expanded and divided into three more levels. According to the User Interface Design Environment (UIDE) (Sukaviriya & Foley, 1993) guidelines for task representation, the actions can be viewed on the high-level application action (e.g. open file), the mid-level interface action (e.g. clicking NEXT), and lastly the low-level interface technique, a mouse click.

Additionally, as mentioned, the educational data can be formatted in several different ways, which also depends on the digital educational environment used. Among others, educational systems are able to calculate test scores, store handed-in digital assignments or student information, and record student system (interface) interactions. These system interactions are nowadays often logged in combination with time information. This temporal data allows for process data analysis, where researchers or teachers able to know when and what a student did in time exactly. If we chronically order this process data and combine it with the structure of the system, we have navigational data, the sequence of actions taken by the student.

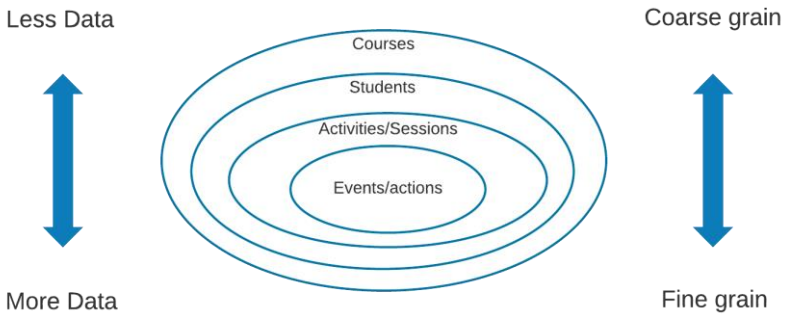


Figure 9 *Relationship between the amount of data and the level of granularity (Romero & Ventura, 2020)*

After we learned the differences in educational system structures and found more levels of hierarchy with the help of UIDE guidelines, we adapted Romero & Ventura's framework (2010) to fit test-taking navigational data and help us in our data exploration (Figure 10).

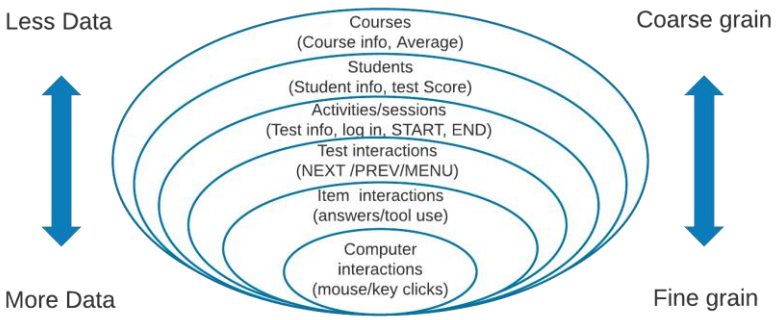


Figure 10 *Adapted relationship between the amount of data and the level of granularity for navigational test-taking*

2.3.2. Educational Data Mining Methods

The application of data mining techniques to the educational domain has become an active research field because of this abundance of data. This is not surprising as there are several applications and tasks that can be resolved with the help of data mining. The most popular and oldest application in education is predicting student performance. However, in recent years, data mining methods have been applied to address more new and different problems. Romero & Ventura (2013) analysed more than 300 studies and identified eleven different categories of EDM applications. With the help of task categorization (overview in Appendix B), researchers (Baker & Yacef, 2009; Bakhshinategh et al., 2018; Bousbia & Belamri, 2014; Castro Espinoza et al., 2007; Scheuer & Mclaren, 2012) identified the data mining methods that can be used to meet the needs of these applications. This categorization is particularly useful for new research, just like ours, as this gives an overview of previous works for a particular task and presents data mining possibilities.

Besides applications, a wide variety of (educational) data mining methods exists and based on the research by Romero & Ventura (2010), the most commonly applied data mining tasks are regression, clustering, classification and association rule mining. Additionally, a complete overview of EDM applications and their most used data mining methods can be found in this research. However, to not go beyond the navigational scope of our research, we will discuss the EDM tasks and methods with the most resemblance to our research's objective.

The first set of applications, predicting students attributes, has the objective to estimate the unknown value of a variable that describes the student (Bakhshinategh et al., 2018). As was mentioned before, the majority of existing EDM studies are focused on this task and try to predict the student's performance, knowledge, score and even engagement or collaboration. Regression and classification were the most widely used methods for this task, but other techniques, such as clustering, have been used as well. Regression analysis finds the relationship between a dependent variable and one or more independent variables (Draper & Smith, 1998). In classification, objects are grouped based on quantitative information regarding one or more characteristics and based on a training set of previously labelled items (Espejo et al., 2010). The decision tree algorithm, which can be either a classification or regression model, is used in several studies involving the predicting task. A decision tree algorithm was used in (Bravo & Ortigosa, 2009) to detect any potential indicators of low performance in e-learning courses.

The second application this research wants to highlight is the detection of undesirable student behaviour. Undesirable is broadly defined and could indicate behaviour such as low motivation, erroneous actions, cheating or dropping out. Most researchers in this category utilize data mining methods like classification, clustering and outlier detection. Clustering identifies groups of objects that are similar (Romero & Ventura, 2013). Unlike classification, the groups do no use

predefined labels and typically use some kind of distance measure to decide how similar objects are. With outlier detection, researchers try to discover data points that are significantly different from the rest of the data. Clustering and outlier detection in the educational domain is often done with the k-means algorithm as it is one of the simplest algorithms to utilize (Dutt et al., 2017) and useful for data exploration without hypothesis (Baker, 2010). The decision tree algorithm was used in the research of Mühlenbrock (2005) to detect anomalies in the learner's actions in web-based learning environments. Clustering with the help of the Kohonen network to detect students cheating in online assessment was done by Burlak et al. (2006). The studies of Nagaoka & Ueno (2004) and Vee (2006) used outlier detection in combination with event logs to detect atypical student behaviour.

Additionally, this research addresses the task of grouping students. Profiling students can be done on various student attributes and can be used to build better personalized and adaptive learning systems (Romero & Ventura, 2010). The data mining often used for this task are classification and clustering, however similar to the other categories, other techniques are applicable too. K-nearest neighbour classification was used to identify student learning styles (Chang et al., 2009). The research of Ayers et al. (2009) used and compared several clustering algorithms, like k-means, to group student with similar skill profiles. Additionally, several studies grouped students based on their interactions with the digital educational environments and interaction pattern. Fok et al. (2005) used a hidden Markov-model-based classification approach, Harley et al. (2013) used an Expectation-Maximization clustering algorithm, and Kinnebrew et al. (2013) utilized a sequence mining technique.

Lastly, we shortly review the visualization task. The goal of the visualization of data is to highlight useful information and support decision making (Romero & Ventura, 2010). The data mining method process mining is especially useful for visualising temporal and navigational data. The goal of process mining is to extract knowledge from event logs recorded by an information system to visualise a particular process (Cristóbal Romero & Ventura, 2013). Through the use of process mining in the educational domain, researchers are able to discover the browsing patterns of students with visualization tools, like ProM[7] and Disco[8].

[7] https://www.promtools.org/doku.php
[8] https://fluxicon.com/disco/

Chapter 3

Methods

This chapter will discuss the data understanding and preparation steps in the CRISP-DM methodology by examining the data, its context and preparation needs. We will describe the selected data and how we created our measures. Afterwards, we will explain how we will answer the sub research questions subject to data analysis.

3.1. Data understanding

The analyses for this research were based on the 'Adaptieve Centrale Eindtoets' (ACET) of 2018. ACET is the Dutch adaptive central final test for students in their last year of primary school. Mandated by the government, The board of Tests and Examinations (College voor Toetsen en Examens, or CvTE) developed this test in collaboration with Stichting Cito. Most students in the final year of primary education in the Netherlands participate in this test to get advice on the type of secondary education to attend. They are tested in the field of reading, grammar & spelling, writing, math and optionally world orientation.

During three or optional four (for world orientation) days, students take different tests, where every test contains 1 to 4 of the subjects, and every subject has 23 items on average. These ACET tests are slightly different from adaptive tests as they have multistage structures. In a multistage test, not items but groups of items are interactively selected for each student. If the student performs well on a particular subject in one stage, he or she will be presented with a more difficult stage in the following test (Figure 11). In most adaptive tests, it is not possible to return to previous questions, but multistage tests, like ACET 2018, allow for students to navigate both forward and backwards to questions within a stage or test.

A total of 22.295 students participated, and 68.204 tests were made, giving this research a good amount of data to analyse.

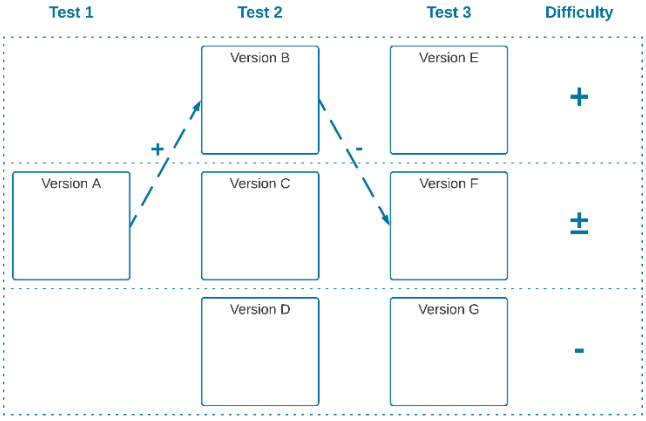


Figure 11 *The Multistage Testing Structure in ACET 2018*

3.2. Data Preparation

The raw data was presented in XML format, a flexible markup language that provides a mechanism to create rules for the storage layout and logical structure (Bray et al., 2000). Figure 12 portrays the structure of the ACET 2018 dataset and Appendix C gives an overview of the description of the data attributes. We made three corrections to the data set during data preparation. Negative and abnormal high (more than 15 minutes) 'ResponseTime' per item were replaced by the average response time for that particular item measured on all students. This accounted for 0.7% of the total responses in the dataset. Secondly, 163 unfinished tests, labelled with the test status 'ABRUPTED', were removed. Lastly, we deleted 632 tests with a total time longer than 4 hours (*mean*= 1.374h, *std*= 0.682h), leaving us with 21.534 students and 67.409 tests.

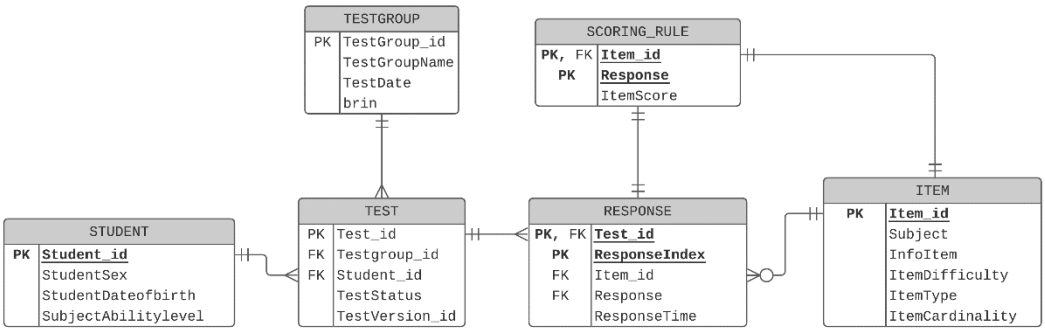


Figure 12 *The Structure of the ACET 2018 dataset*

An attribute this research wants to highlight is the 'ResponseIndex', which shows the order of actions of a student. This variable alone does not say much about the navigation sequence since it is the cumulative sum of all steps. Moving from item 1 to item 2 back to item 1 would result in a 'ResponseIndex' of 1, 2, 3. Additional item order information was missing in the original test design. Therefore, we had to derive and construct the order of the items in the test, which is the cumulative count of unique items seen by the student.

With this item presentation order, we were able to get the distance between items and discover how the students moved through the test. With the help of this and measure and the literature study, we distinguished seven activities in our data (Table 4).

Table 4

Test-taking navigation activities distinguished in the data

Test-taking navigation activities	Description
Forward	Moving forward by clicking the next button or clicking the next item
Forward Jump	Moving forward, but not to the direct next item
Backwards	Moving backwards by clicking the previous button or clicking the previous item
Backwards Jump	Moving backwards, but not to the previous direct item
Skip	Moving forward or backward without giving a response to an item
Change	Moving forward or backwards and changing the previous response to a different response
Item hop	Moving forward or backwards within 5 seconds without changing the answer
Info hop	Moving forward or backward from an info item within 5 seconds

The first four activities are solely based on the distance between items, but the other activities need the help of more than one attribute. *Changing* and *skipping* requires the student's response to an item. For *skipping,* we also needed to distinguish the difference between question items and information items. Information items give instruction for the next set of question items and do not require a response. *Hopping* was added after we found such activities in our preliminary data analysis in the Disco visualization tool. *Hopping* is based on both navigational and temporal data. We assumed that a student is very unlikely to answer an item within 5 seconds and might use a *hop* in order to navigate quickly to different items. In *hopping,* we distinguish between question items and information items too. More *hopping* on information items in a later stage of the test could confirm the theory of *hopping* as a navigating strategy.

Lee & Haberman (2016) and Vasilyeva et al. (2010) discovered that students navigate more flexibly towards the end of high-stake tests. We identified such behaviour in the navigation paths of our students (see Appendix D) through our preliminary data analysis in Disco. However, several students changed their behaviour entirely after the last question, possibly finishing the test and checking the answers on previous (found difficult) items. Some students did not act after the last question. Therefore, we separated the navigation behaviour into two categories: *before-last-item* and *after-last-item* behaviour. Navigation behaviour in these two categories could be very different and probably has a different motivation.

In order to compare the students' navigation behaviour, we divided the frequency of the activities by the number of unique items per test, as the number of items differed per test. Finally, we got the navigation behaviour profile per student by averaging the activity attributes over the number of tests taken.

3.3. Method

For research question 3.2, we tried to identify common navigation strategies in tests. We used the activities (Table 4) on the levels *before-last-item* and *after-last-item* and on the combination of those to group the students. Next, we scaled the frequency of the activities with a min-max normalization. Afterwards, we reduced the dimensionality of the multivariate structure of our data with Principal Component Analysis and found the optimal number of clusters. The number of clusters were used in the k-means clustering algorithm and were finally plotted in a grouped bar chart. A snippet of the code in python and an example graph showing the optimal clusters can be found in Appendix E. Additionally, as we discovered in the literature (Schnipke & Scrams, 1997) that rapid-guessing behaviour was a function of item position, we were interested if navigation behaviour was also influenced by item position. Therefore, for the last level, we divided the activities not into the two groups *before* and *after*, but in four groups *early before*, *mid before, late before* and finally *after,* based on the item position in the test.

Research question 4.1 involved analysing the relation between navigational behaviour and the student's performance. For student's performance, we took the estimated overall ability of a student, as individual test scores are not sufficient due to the adaptive multistage tests. Recall that better performers get more difficult tests the next day. The estimated overall ability is calculated by combining the plausible values of reading, writing and math. Plausible values are based on student responses, as well as on other relevant information, like background information (Von Davier et al., 2009). We did not take students who only took world orientation in this calculation, as their plausible value would be zero. Afterwards, we evaluated if the mean of the estimated student ability differed between the clusters of navigation behaviour found in question 4.1. Additionally, we analysed if and how the navigation activities relate to student ability.

In research question 4.2, we mimicked the research of Jacobs (1972) and analysed if students changing their answer is for the better or worse. We also examined if less difficult items were more correctly revised than more difficult items.

In the last research question, question 5, we continued to look for differences between (non-info) items and navigation activities. Additionally, as we found that rapid-guessing behaviour was a function of item position (Schnipke & Scrams, 1997), we too were interested if there was such a relationship between navigation behaviour and item position. We analysed item position in terms of percentages and divided the previous item order into 50 bins of 2%. This option was chosen as most tests had less than 100 items. Lastly, we analysed the navigation activities per item. We were curious if outlier items, e.g. items with more proportional more skipping, also have a higher predetermined item difficulty. If outlier items do not have higher item difficulty, it might be interesting to investigate that item. What made students skip that item more?

Chapter 4

Results

This chapter will evaluate the findings after performing the data analysis and modelling. We will use the sub research questions as a guide on discussing the results.

4.1. Overall

We analysed the navigation paths of 21.534 students. The mean path length was 1.43 times the length of the test, with an average response time of 43.46 seconds. 86% of the navigation actions were moving to the next item in a linear fashion. The average distance of steps was 0.65, indicating that students moved backwards with higher jumps. Table 5 gives an overview of the total descriptive statistics.

Table 5

Overview of the total descriptive statistics on student's navigation behaviour

	Path length	Distance steps	Linearity	Empty items	Response time (in s)	Total test time (in h)
mean	1.43	0.65	0.86	0.00	43.46	1.36
std	0.39	0.21	0.10	0.02	11.70	0.47
min	1.00	-0.24	0.33	0.00	10.21	0.28
25%	1.14	0.50	0.81	0.00	35.25	1.04
50%	1.29	0.67	0.88	0.00	42.06	1.26
75%	1.56	0.83	0.93	0.00	50.10	1.56
max	5.07	0.99	0.99	0.28	130.00	3.96

The first behaviour that could indicate a navigation strategy was found whilst examining path linearity. In nearly 3% of the tests, no other navigation actions besides forwarding occurred, and in 0.3% of tests, every item was seen precisely twice. We analysed whether the students that only visit an item once in linear order do this because they are out of time, but this was not the case (Figure 13). Mean test time was 1 hour, and only one test had more than the recommended 3 hour test time.

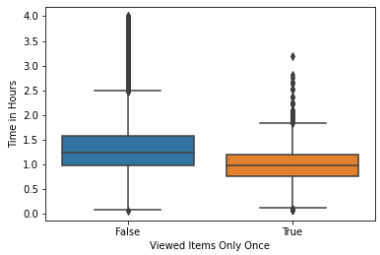


Figure 13 *Time spent on test grouped by students only viewing item once or more*

4.2. Group

For research question 3, we tried to discover which navigational behaviour exists in our data and if there is a possibility to distinguish between different strategies. With the help of the process data visualization tool Disco, we discovered that not all students navigate after completing the last item, potentially indicating a difference in navigational behaviour. With data analysis, we were able to confirm that 23.74% of the students did not take a step after the last item. Additionally, we analysed all the activities and distinguished between before and after last item behaviour. One finding was that 99.71% of the students change at least one response in the test and change on average 11.9% of the items per test. Furthermore, we discovered that most students will mainly use the standard forwards (80.65% of the times), skips (11.9%) and item hops (19.4%). An overview of the statistics on the activities can be found in Figure 14.

No test-taking navigational differences were found in the literature study. In combination with the multivariate character of the data, this makes it hard to understand what we are looking for. This research tried finding navigational differences with the help of the k-means clustering algorithm. On the three levels, *before-last-item*, *after-last-item*, and total, we first identified the optimal number of clusters and later ran the algorithm on the navigation activities (Table 4).

	Mean
total forward	0.85
before forward	0.80
after forward	0.05
total forward jumps	0.07
before forward jumps	0.05
after forward jumps	0.02
total backs	0.08
before backs	0.07
after backs	0.01
total back jumps	0.06
before back jumps	0.04
after back jumps	0.01
total skip	0.11
before skip	0.11
after skip	0.00
total changed	0.11
before changed	0.10
after changed	0.01
total item hops	0.19
before item hops	0.11
after item hops	0.08
total info hops	0.09
before info hops	0.08
after info hops	0.01

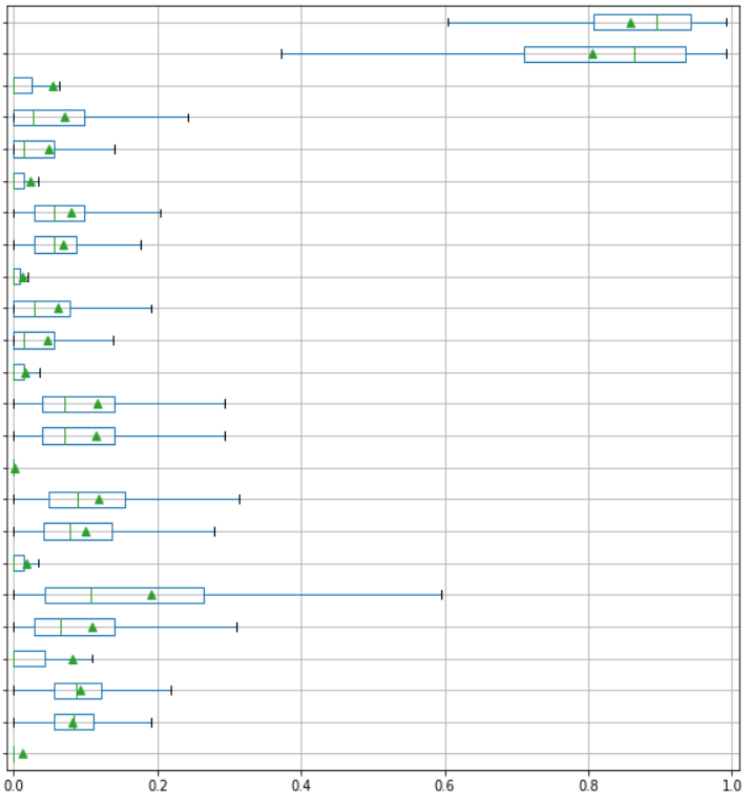


Figure 14 *The distributions and averages of the navigation activity occurrences. (Green triangle = mean)*

For the *before-last-item* navigation behaviour, we identified the optimal number of two clusters. The two distinguished clusters are shown by a grouped bar chart in

Figure 15. 69.5% of the students fell into the purple cluster. Most students in this group move mostly linearly until the end of the test. The smaller, orange group can be identified by more activity on all other activities and skip or hop considerably more items.

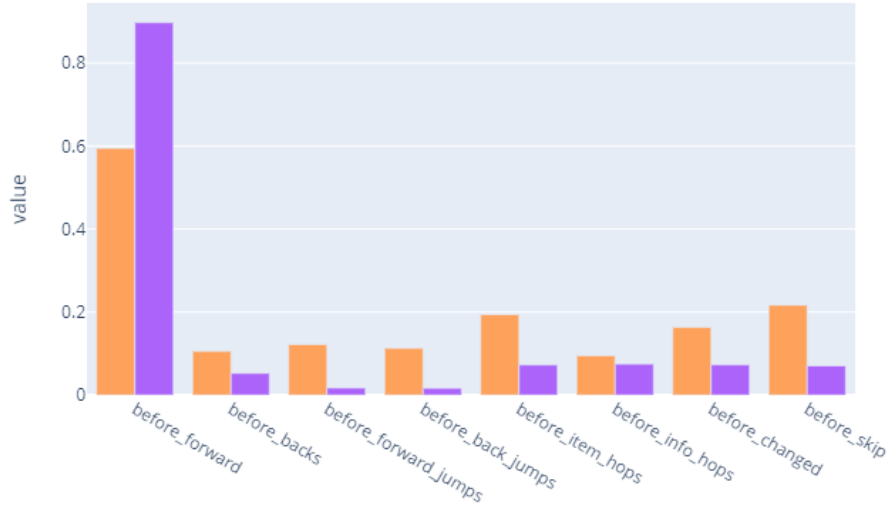


Figure 15 *Groups of students clustered using before-last-item navigation measures (*purple *= mostly linear,* orange *= more activity)*

Navigational behaviour *after the last item* was clustered into two groups as well using the same k-means clustering algorithm. As one can see in Figure 16, the yellow group, which is actually the vast majority of students (88.8%), has little activity after the last item has been made. The other 11.2% of the students (pink) do navigate after the last item and uses the quick *hops* and long *jumps* to probably navigate to check items that were marked and found difficult.

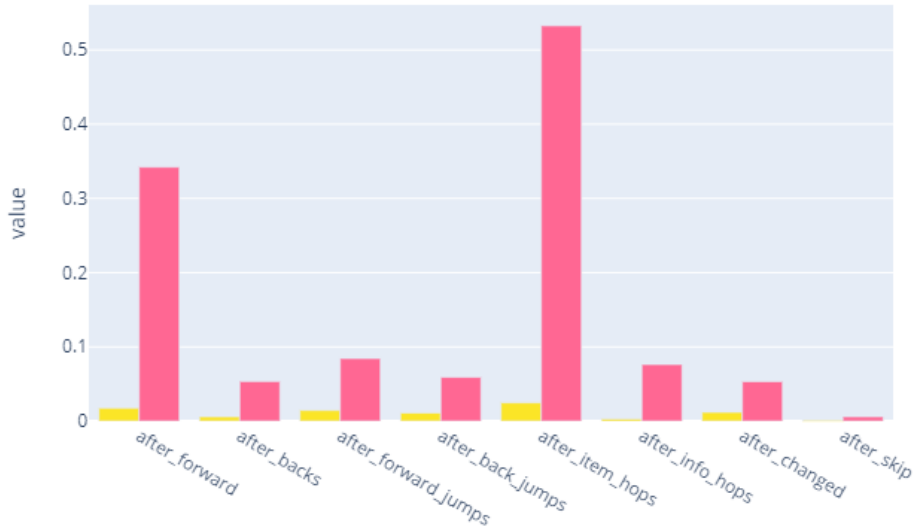


Figure 16 *Groups of students clustered using after-last-item navigation measures (*yellow *= no activity,* pink *= more activity)*

Lastly, we combined the *before* and *after* activities and ran the algorithm once again to identify groups of students on all the activities in a test. This time the algorithm identified three clusters: the fairly linear group, the more movement group and lastly, the group of the students that acts after the last item in the test. 69.4% of the students were selected for the green group, which mainly moved

linearly. The second-largest group was the red one (17.2%), with more *item hopping*, *item changes* and *skipped items* before reaching the last item. The blue group, accounting for 11.4% of the total students, showed increasingly more action after reaching the last item than the two other groups. The bar graph in Figure 17 represents the clusters found at this level.

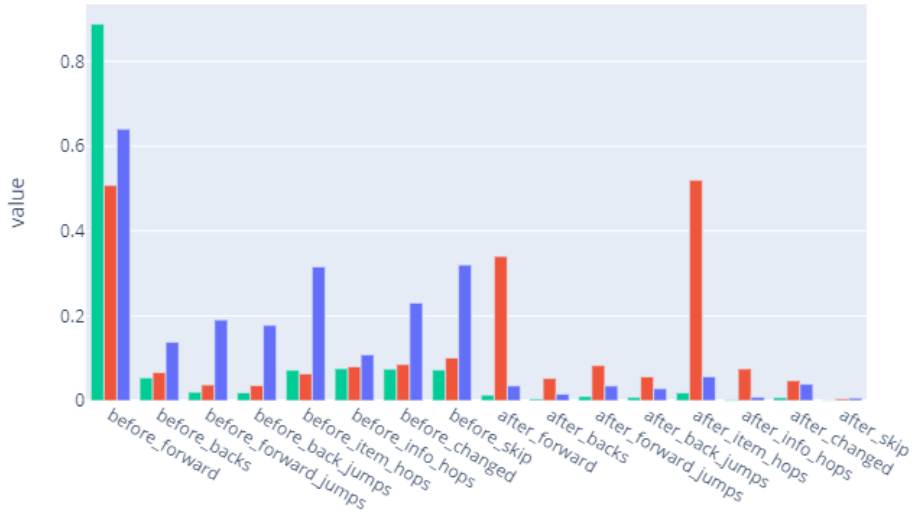


Figure 17 *Groups of students clustered using all navigation measures (*blue *= mostly linear,* red *= more activity before last item,* green *= more activity after last item)*

The last level for clustering added to the previous found clusters and provided additional information by looking at item position. The graph, shown in Appendix F with *before* activities divided into the three groups early (1), mid (2) and late (3) item positions, shows the same three clusters as the previous level. This analysis gives us especially more detail about the red, more movement, group. Students in this group move less linearly in the later sections than in the first early section. Additionally, the red group skips more items in the first two sections and change them in the mid and late section.

Furthermore, as our data had multiple tests per student, we analysed whether students had consistent navigational behaviour on all tests, possibly indicating a preferred test-taking style. We utilized the clusters found on all navigational activities. Of all 21.534 students, a little more than half (54.1%) was consistently clustered, and 38.2% of the students had only one test where they navigated differently according to the clusters (Appendix G).

4.3. Student

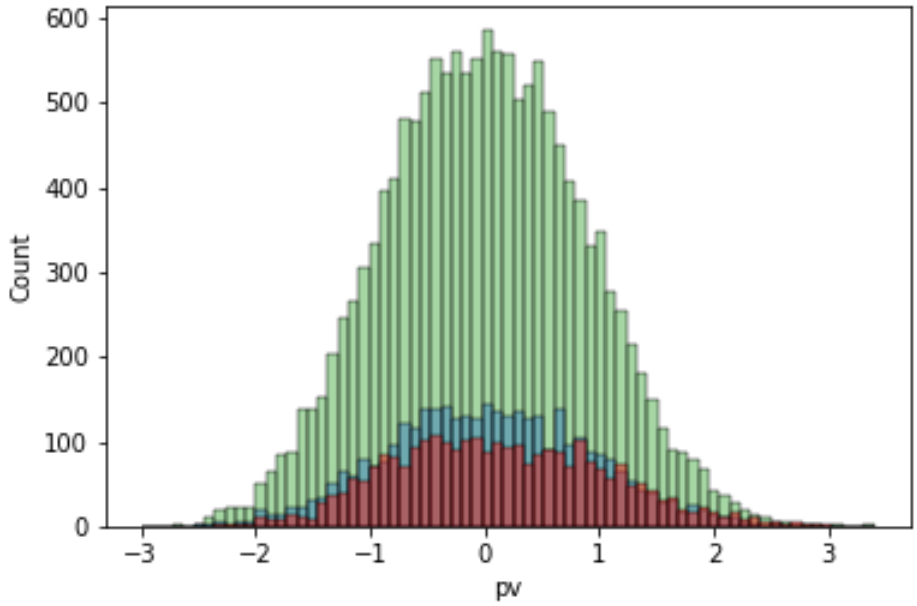
There was no strong relation between a student's ability and any of the navigation activities. The activities with the greatest, however, still likely to be unimportant, positive correlation was before backs ($r_{before\ backs}$=0.105). The most significant negative correlation was -0.138 with the activity measure total distance (Table 6).

Table 6

Correlation between navigation measures and student ability

Navigation measure	Correlation
before backs	0.105
total steps	0.099
total backs	0.096
before steps	0.089
total info hops	0.075
after forward	0.068
after steps	0.066
before info hops	0.058
after info hops	0.052
after back jumps	0.049
after item hops	0.045
total item hops	0.038
total changed	0.034
after changed	0.032
after backs	0.029
before changed	0.022
total back jumps	0.017
after forward jumps	0.016
total skip	0.013
before skip	0.013
after skip	0.011
total forward jumps	0.008
before item hops	0.003
before back jumps	0.001
before forward jumps	0.001
total empty	-0.029
total forward	-0.031
after distance	-0.038
before forward	-0.071
before distance	-0.102
total distance	-0.138

Additionally, we examined the ability level differences in the previously found clusters (Figure 18). On the total navigation behaviour level, we discovered a significant difference between the means of the groups. (*One-way ANOVA test* ($_F$=32.73, $_{p\ value}$=6.44e-15)), which is presumably the effect of the sufficiently large sample size. P-value can inform whether an effect exists, but it will not reveal the size of the effect (Sullivan & Feinn, 2012). For this reason, we used Cohen's effect size measure. According to Cohen (1988), the effect size is low if the value is around 0.1, is medium if it varies around 0.3, and is large if more than 0.5. It was found that the differences between our groups' ability mean are negligible as the calculated effect sizes are 0.04 (green-blue), 0.15 (green-red) and 0.11 (blue-red).



	count	%	mean	std
Mostly linear (green)	14575	69.4	-0.0005	0.874
More active (blue)	3591	17.1	0.034	0.880
Active after (red)	2825	13.5	0.138	0.946

Figure 18 *The statistics on student performance by clusters on all navigation measures*

SSQ 4.2 was concerned with the impact of the navigational activities on the correctness of a single item. Therefore, we calculated the mean scores of the items on all students and looked for a correlation between these scores and the navigational activities. The resulting correlations can be found in Table 7. Skipping an item had a relatively small negative correlation with item score ($r_{total\ skips}$=-0.232. Hopping an item after the last item of a test was one of the few activities with a positive, though insignificant, correlation to item score ($r_{after\ hops}$=0.152). However, hopping and especially hopping before the last item of a test ($r_{before\ hops}$=-0.303) had a higher negative correlation. The activity with the highest negative correlation was changing an answer of an item ($r_{total\ changed}$=-0.319).

Table 7

Correlation between navigation measures and item score

Navigation measure	Correlation
after hops	0.152
before distance	0.050
after forward	0.027
after back jumps	0.006
after steps	-0.012
total forward	-0.032
after forward jumps	-0.038
total distance	-0.039
before forward jumps	-0.049
total forward jumps	-0.055
after distance	-0.064
before forward	-0.065
before mean time	-0.074
before backs	-0.090
total back jumps	-0.101
before empty	-0.108
total steps	-0.110
total empty	-0.110
after empty	-0.110
after backs	-0.121
before steps	-0.122
total backs	-0.140
before back jumps	-0.155
after changed	-0.213
before skips	-0.216
after skips	-0.216
total skips	-0.231
total hops	-0.243
before changed	-0.247
before hops	-0.302
total changed	-0.318

Next, we replicated the answer changes research and table of Stanley Jacobs (1972) and discovered even more positive results than found previously. Of the total answer changes, 59.43% was changed from an incorrect response to a correct response (Table 8). Additionally, we found no discernible differences in answer changes among the different clusters found on the total activity level (Appendix H). The most significant difference with Stanley Jacobs' research was the much lower total percentage for incorrectly correcting an item, 3.7% instead of 20.1%.

Table 8

Types of Answer Changes Made to Items of different difficulties in our dataset

(n=21.534 tests, 628.446 changes)	Item difficulty			
	Low	*Moderate*	*High*	Total
Right to Wrong	1.0%	1.2%	1.5%	3.7%
Wrong to Right	21.2%	20.1%	18.1%	59.4%
Wrong to Wrong	7.9%	11.9%	17.1%	36.9%
Total	30.1%	33.2%	36.7%	100%

4.4. Items

The final research question continued the exploration of the relation between items and navigational activities. As we can see in
Table ***8*** above, difficult items are changed relatively more often, and easy items are more often changed for the better. Additionally, we analysed the relations between the navigational activities and the predetermined item difficulty. As item difficulty and item score relate reversely, the results were found to follow this as well (Table 9).

Table 9

Correlation between navigation measures and item difficulty

Navigation measure	Correlation
after hops	-0.036
before forward jumps	-0.012
before back jumps	0.008
after back jumps	0.012
total back jumps	0.012
total hops	0.030
before distance	0.042
total forward jumps	0.052
before hops	0.061
before forward	0.068
after distance	0.069
total distance	0.071
before steps	0.082
total forward	0.107
after forward	0.113
after forward jumps	0.131
before skips	0.131
total steps	0.144
total skips	0.147
after empty	0.150
after steps	0.154
before changed	0.165
before backs	0.168
after backs	0.170
total empty	0.211
before empty	0.214
total backs	0.239
total changed	0.269
after changed	0.270
after skips	0.310

Table 10

Correlation between navigation measures and item presentation position

Navigation measure	Correlation
after distance	0.979
after forward jumps	0.675
before empty	0.647
total empty	0.643
after backs	0.586
after empty	0.579
total forward jumps	0.562
after skips	0.526
after changed	0.444
before distance	0.355
before forward jumps	0.260
after steps	0.259
after hops	0.249
after forward	0.125
total forward	-0.084
total hops	-0.085
before forward	-0.126
total steps	-0.153
total changed	-0.176
total score	-0.217
before steps	-0.246
total skips	-0.252
before skips	-0.283
before hops	-0.344
before changed	-0.354
after back jumps	-0.354
total backs	-0.459
before backs	-0.643
total back jumps	-0.711
before back jumps	-0.823

Most activity measures had a positive correlation with item difficulty. Additionally, answer changes and item skipping after the last item had the highest correlation with item difficulty ($r_{after\ changed}$=0.270, $r_{after\ skip}$=0.310). In contrast to item score, this analysis showed a more significant correlation with item backing ($r_{total\ backs}$=0.239), and item hopping had almost no correlation anymore; from $r_{total\ hops}$ =-0.243 to $r_{total\ hops}$ =0.003.

The following analysis was concerned with the analysis of item presentation position. Table 10 gives the user a complete overview of the correlation between

item position and navigational activities. All types of backing activities, except backing after the last item, had a high negative correlation with item position. This is obvious as the student is never able to back to an item that has the maximum value in item presentation order. Later items were less likely to be skipped, hopped and changed, but were more likely to be left empty before the last item of the test. However, late items were usually reviewed after the end of the last question with more steps, including backing and jumping afterwards.

For our last analysis, we briefly studied the items with an outlier frequency of navigational activities. We took skipping before the last item as an example. Items were, on average, skipped in 12% of the tests. However, there were 25 out of the 600 non-info items with higher skipping percentages, portrayed as the dots above the boxplots in Figure 19. 7 of the outlier items had predetermined item difficulty lower than the average item difficulty (=-0.737).

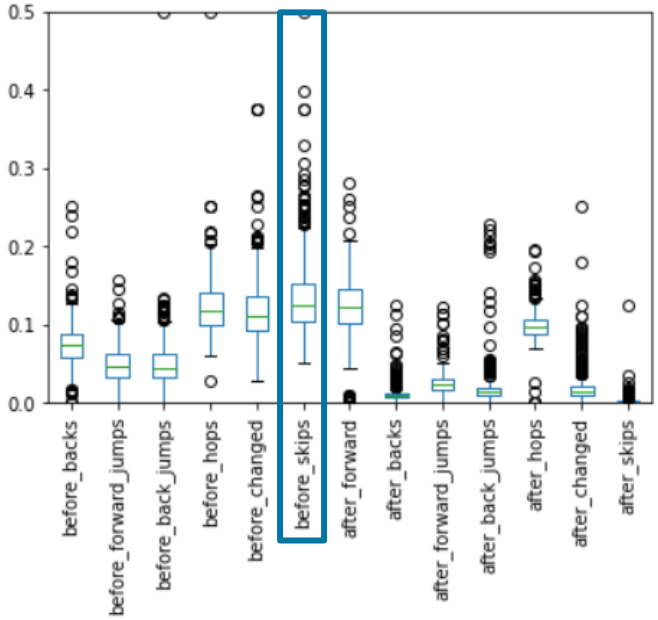


Figure 19 *Boxplot of Navigation Activities per Items*

Chapter 5

Conclusion

The digital shift in the educational test-taking sector has led to the availability of more data. Where previously this domain was focussed on performance and score, with the help of more data, nowadays, researchers are analysing how and why students are interacting with educational matter. This research extends the exploration of this data and addresses the analysis of navigational test data.

SQ1 *What different test-taking navigational behaviour exist in the literature?*

In order to understand how and why students navigate through a digital educational system, it is crucial to identify the student's task, the student's attributes, e.g., expertise and cognitive style and finally, the structure of the system and its navigational tools. Differences in these factors exist and alter navigation behaviour. Studies in web and EHS navigation behaviour (Bousbia et al., 2010; Canter et al., 1985) are leading in the exploration of navigational strategies and proposed typologies for 'hypermedia' navigation types. Mainly due to the lower dimensionality of a test and the relatively more limited navigational tools, one cannot simply compare test-taking navigation behaviour to other digital navigation behaviour, not even within the educational domain. In addition, studies in test-taking navigation behaviour are scarce, and there are no strategies to be discovered. However, studies in test-taking behaviour do mention students will mostly follow the linear order of the test (Lee & Haberman, 2016; Vasilyeva et al., 2010) and advises analysing the out-of-sequence anomalies, like *jumping*, *skipping* and *changing* a question.

SQ2 *What techniques can be used for inferring navigational behaviour from digital tests?*

Educational data mining applications and techniques have been analysed and categorized by researchers. The most commonly applied data mining tasks are regression, clustering, classification and association rule mining. The most popular and oldest application for educational data mining methods is predicting the student's performances. However, in recent years, data mining methods have been applied to address more new and different problems. This research identified three application tasks similar to our research objectives and within the test-taking navigation scope. We analysed educational studies grouping student navigation patterns and found that the k-means clustering is commonly used for this application in the education domain. Process data and navigational patterns were often analysed with process mining and data visualization. Therefore, we used these techniques to preliminary analyse the data.

After a preliminary data analysis, we devised a set of frequently found navigation activities. Besides the linear *steps* (distance =+1/=-1), we added deviating activities with out-of-linear *jumps* (>= +2/<= -2) *forwards* and *backwards*, item *skips*, meaning unanswered questions and item *changes*, where the previous answer is changed by the student. Lastly, we distinguished the last activity *hopping*, navigating to another item within 5 seconds. Additionally, it turns out to be helpful to distinguish navigation *before* the end of a test and *after* as one is able to capture a more accurate representation of test-taking navigation behaviour.

SQ3 *What type of navigational behaviour can we discover in our data?*

Typically, students use more activities than the number of items in a test. As to be expected, the most common navigation activity was *forwarding*, occurring 85.9% of the times. However, the average total distance was 0.66, indicating that students used more or larger backward jumps than backs when moving out of the linear navigation order. There is a small group of students (2.9%) that only visit an item once in linear order and submit the test after the last item. Most of them had time left to check answer but chose not to. Future research should consider examining these students and their reasons not to navigate differently.

With the k-means clustering, we identified three test-taking navigation strategies. The largest group, good for 66.3% of the students, is the mostly linear group. 21.2% of the students navigate more than the previous, with more item hopping, item changes and skipped items before reaching the last item. The final group (12.6%) showed increasingly more activity after reaching the last item. In addition, we discovered that students showed the same navigation behaviour (strategy) on all the different tests. A little more than half of the students had the same navigation strategy on all the test, and 38.2% of the students had only one test where they navigated differently according to the clusters. Although the identified strategies and consistency might not differ in performance, it could help us understand the cognitive attributes and preferences of a student.

SQ4 *How does navigational test behaviour relate to the performance of students?*

It appears there is little relation between the navigational activities and student ability. Additionally, the different navigation strategies discovered by the clusters had no differences in student ability as well. On item performance level, we discovered that the number of answer changes (average of 11.9% of the items) is negatively related to item score ($r_{total\ changed}$=-0.319). However, the correlation needs further experiment with controlled variables to understand the causation effects. Is the answer of the item wrong due to the number of changes, or are items with a higher score less changed? The same is true for the navigational activity total skipping ($r_{total\ skips}$=-0.231), possibly indicating that students skip items that they found difficult.

Besides the number of answer changes, this research zoomed in on the impact of a single answer change on the item score. Changing from an incorrect response to a correct response happened 59.43% of the total answer changes. Only 3.68% of the total answer changes were disadvantageous, which indicates that reconsidering an item could be worth it. This finding helps to advocate the use of free navigation in tests. Students might not know the answer at that moment and need an extra chance to show their full potential.

SQ5 *How does navigational test behaviour relate to a test and its items?*

Skipping and changes of answers occur most frequently with difficult items ($r_{after\ changed}$=0.270, $r_{after\ skip}$=0.310). Items with a higher number of non-linear activities and low item difficulty or item score need further investigation. The non-linear activities could serve as an indicator for difficult items and might be beneficial for the teacher and students to recapitulate during feedback. Perhaps test makers could use the navigation metrics to improve the item difficulty function or find hard to understand questions.

In our last analysis on item position, we found that later items were less likely to be skipped, hopped and changed, but were more likely to be left empty before the last item of the test. However, late items were usually reviewed after the end of the last question with more steps, including backing and jumping afterwards.

Chapter 6

Discussion

6.1. Limitations

There are a number of limitations in our research that should be considered as caveats. First, it should be considered that this navigational analysis was done on primary school students. Age and school level should be taken into account when comparing results with this research as it might impact the navigational behaviour of students.

The first limitation of the research is in the structure of the ACET 2018 test. Within one test, a student is often presented with multiple, often dissimilar subjects per test. On average, a test had three blocks of 23 items with the same subject. Although this research does not know how and if this impacts the results presented, we hypothesize differences with “one subject” tests. We expect that students are less likely to go back and review items dissimilar to the subject of the current item. Additionally, we expect differences with “after last item” behaviour.

Secondly, we made some assumptions due to the not easily available predetermined item positions per test version. Although this test had an adaptive structure, the adaptiveness was between tests and not items. There were several predetermined versions of blocks of items with different item orders. Unfortunately, we were unable to retrieve the block versions or item orders timely, and therefore we had to construct our own order. With the help of the literature, we found and assumed that most students will start with the first item of the test. We created an item order function that cumulatively counts when a new unique item is seen. However, this is not representative as students could potentially forward jump from the first item to the third. Our function will wrongfully label this item as the second item in the test. This could give significant differences in results. Future research with the ACET 2018 dataset should try to retrieve and use the predetermined item order.

6.2. Future Research

In addition to the limitations, there are several opportunities that can potentially benefit future research. Future research should consider adding navigational interactions, like menu clicks and marking items. A system may permit a student to flag/mark the item to review an (unanswered) item that needs some more attention at a later time. Researchers could analyse the correlation between marking items and “after last item” behaviour and if marking is beneficial for the test score. Besides analysing the between-item navigation behaviour, like this research, researchers could dive deeper into in-item navigation behaviour. Clicking, keyboard strokes and tool usage might reveal individual test item strategies.

The choice of clustering algorithm could merit further investigation. This research took the k-means clustering algorithm as it is often used in the educational domain. However, it might be interesting to analyse and compare the findings of several different algorithms.

The biggest opportunity is contextualizing the correlations and findings. Although the limited research in navigational test-taking behaviour, future research should try to seek domain experts that could interpret the navigational behaviours, the differences and findings. By the ACET 2018 data set and our bottom-up data analysis approach, we are only able to report the correlations between navigational activities and other attributes. Future research could design experiments with controlled variables to analyse possible causation effects. Researchers could identify interesting results by knowing the student level of ability, cognitive style or navigational preferences prior to the test.

References

Akçapınar, G., & Altun, A. (2010). The effect of prior knowledge on learners' navigation structure. *Proceedings of the IADIS International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2010*, 273–276.

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics Informatics*, *37*, 13–49.

ALTY, J. L. (1984). The application of path algebras to interactive dialogue design. *Behaviour & Information Technology*, *3*(2), 119–132. https://doi.org/10.1080/01449298408901743

Antonietti, A., & Giorgetti, M. (1998). The Verbalizer-Visualizer Questionnaire: A Review. *Perceptual and Motor Skills*, *86*(1), 227–239. https://doi.org/10.2466/pms.1998.86.1.227

Ayers, E., Nugent, R., & Dean, N. (2009). A Comparison of Student Skill Knowledge Estimates. *EDM*.

Baker, R. (2010). Data mining for education. *International Encyclopedia of Education*, *7*(3), 112–118.

Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*, 3–17.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*(1), 537–553. https://doi.org/10.1007/s10639-017-9616-z

Barab, S. A., Bowdish, B. E., & Lawless, K. A. (1997). Hypermedia navigation: Profiles of hypermedia users. *Educational Technology Research and Development*, *45*(3), 23–41. https://doi.org/10.1007/BF02299727

Baumgarten, M., Buchner, A. G., Anand, S. S., Mulvenna, M., & Hughes, J. (2000). *User-driven navigation pattern discovery from Internet data*.

Bezirhan, U., von Davier, M., & Grabovsky, I. (2020). Modeling Item Revisit Behavior: The Hierarchical Speed–Accuracy–Revisits Model. *Educational and Psychological Measurement*, *0*(0), 0013164420950556. https://doi.org/10.1177/0013164420950556

Bingman, V. P., Ioalé, P., Casini, G., & Bagnoli, P. (1990). The avian hippocampus: Evidence for a role in the development of the homing pigeon navigational map. In *Behavioral Neuroscience* (Vol. 104, Issue 6, pp. 906–911). American Psychological Association. https://doi.org/10.1037/0735-7044.104.6.906

Bıçak, B. (2013). Scale for Test Preparation and Test Taking Strategies. *Kuram ve Uygulamada Egitim Bilimleri*, *13*, 279–289. https://doi.org/10.1037/t32846-000

Bousbia, N., & Belamri, I. (2014). Which Contribution Does EDM Provide to Computer-Based Learning Environments? In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 3–28). Springer International Publishing. https://doi.org/10.1007/978-3-319-02738-8_1

Bousbia, N., Rebaï, I., Labat, J. M., & Balla, A. (2010). Learners' navigation behavior identification based on trace analysis. *User Modeling and User-Adapted Interaction*, *20*(5), 455–494. https://doi.org/10.1007/s11257-010-9081-5

Bravo, J., & Ortigosa, A. (2009). Detecting Symptoms of Low Performance Using Production Rules. *EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining*, 31–40.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., & others. (2000). *Extensible markup language (XML) 1.0*. W3C recommendation October.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). *Graph structure in the Web*.

Broekkamp, H., & van Hout-Wolters, B. (2007). The gap between educational research and practice: A literature review, symposium, and questionnaire. *Educational Research and Evaluation*, *13*(3), 203–220. https://doi.org/10.1080/13803610701626127

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized educational measurement. *ETS Research Report Series*, *1988*(1), i--148.

Burlak, G., Muñoz-Arteaga, J., & Hernández[1], J.-A. (2006). *Detecting Cheats In Online Student Assessments Using Data Mining.*

Calvet Liñán, L., & Juan Pérez, Á. A. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. *International Journal of Educational Technology in Higher Education*, *12*(3), 98–112. https://doi.org/10.7238/rusc.v12i3.2515

Cambridge University. (n.d.). Behaviour. In *Cambridge Advanced Learner's Dictionary & Thesaurus*. https://dictionary.cambridge.org/dictionary/english/behaviour

Canter, D., Rivers, R., & Storrs, G. (1985). Characterizing user navigation through complex data structures. *Behaviour & Information Technology*, *4*(2), 93–102. https://doi.org/10.1080/01449298508901791

Carr, A. (1965). The Navigation of the Green Turtle. *Scientific American*, *212*(5), 78–87. http://www.jstor.org/stable/24931878

Castro Espinoza, F., Vellido, A., & Nebot, À. (2007). Applying Data Mining Techniques to e-Learning Problems. In *Evolution of Teaching and Learning Paradigms in Intelligent Environment* (Vol. 62, pp. 183–221). https://doi.org/10.1007/978-3-540-71974-8_8

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide web. *Computer Networks and ISDN Systems*, *27*(6), 1065–1073. https://doi.org/https://doi.org/10.1016/0169-7552(95)00043-7

Chang, Y.-C., Kao, W.-Y., Chu, C.-P., & Chiu, C.-H. (2009). A learning style classification mechanism for e-learning. *Computers & Education*, *53*(2), 273–285. https://www.learntechlib.org/p/67058

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. R., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.

Chen, L.-H. (2010). Web-based learning programs: Use by learners with various cognitive styles. *Computers & Education*, *54*(4), 1028–1035. https://doi.org/https://doi.org/10.1016/j.compedu.2009.10.008

Cito. (2021). https://www.cito.com/

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley \& Sons.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, *5*, 15991–16005.

Espejo, P. G., Ventura, S., & Herrera, F. (2010). A Survey on the Application of Genetic Programming to Classification. *Trans. Sys. Man Cyber Part C*, *40*(2), 121–144.

Eveland, W., & Dunwoody, S. (1998). Users and Navigation Patterns of a Science World Wide Web Site for the Public. *Public Understanding of Science*, *7*, 285–311. https://doi.org/10.1088/0963-6625/7/4/003

Felder, R. M., & Silverman, L. K. (1988). Learning and teaching styles in

engineering education. *Journal of Engineering Education*, *78*(8), 674–681.
Ferguson, K. J., Kreiter, C. D., Peterson, M. W., Rowat, J. A., & Elliott, S. T. (2002). Is that your final answer? Relationship of changed answers to overall performance on a computer-based medical school course examination. *Teaching and Learning in Medicine*, *14*(1), 20–23. https://doi.org/10.1207/S15328015TLM1401_6
Fischer, M. R., Herrmann, S., & Kopp, V. (2005). Answering multiple-choice questions in high-stakes medical examinations. *Medical Education*, *39*(9), 890–894. https://doi.org/10.1111/j.1365-2929.2005.02243.x
Fok, A. W. P., Wong, H. S., & Chen, Y. S. (2005). Hidden Markov Model Based Characterization of Content Access Patterns in an e-Learning Environment. *2005 IEEE International Conference on Multimedia and Expo*, 201–204. https://doi.org/10.1109/ICME.2005.1521395
Gall, J. E., & Hannafin, M. J. (1994). A framework for the study of hypertext. *Instructional Science*, *22*(3), 207–232. https://doi.org/10.1007/BF00892243
Graf, S., Liu, T.-C., & Kinshuk. (2010). Analysis of learners' navigational behaviour and their learning styles in an online course. *Journal of Computer Assisted Learning*, *26*(2), 116–131. https://doi.org/https://doi.org/10.1111/j.1365-2729.2009.00336.x
Graff, M. (2005). Individual differences in hypertext browsing strategies. *Behaviour & IT*, *24*, 93–99. https://doi.org/10.1080/01449290512331321848
Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46. https://doi.org/https://doi.org/10.1016/j.chb.2016.02.095
Harley, J. M., Bouchet, F., Trevors, G., & Azevedo, R. (2013). Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning. *EDM 2013*.
Herder, E., & Juvina, I. (2004). Discovery of Individual User Navigation Styles. *Proc. Workshop on Individual Differences in Adaptive Hypermedia, Held at AH2004*.
Herder, E., & van Dijk, B. (2004). *Site Structure and User Navigation: Models, Measures, and Methods*.
Hong, E., Sas, M., & Sas, J. C. (2006). Test-Taking Strategies of High and Low Mathematics Achievers. *The Journal of Educational Research*, *99*(3), 144–155. https://doi.org/10.3200/JOER.99.3.144-155
Hwang, G.-J., Spikol, D., & Li, K.-C. (2018). Guest Editorial. *Journal of Educational Technology & Society*, *21*(2), 134–136. http://www.jstor.org/stable/26388386
Iivari, N., Sharma, S., & Ventä-Olkkonen, L. (2020). Digital transformation of everyday life – How COVID-19 pandemic transformed the basic education of the young generation and why information management research should care? *International Journal of Information Management*, *55*, 102183. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2020.102183
Jacobs, S. S. (1972). Answer Changing On Objective Tests: Some Implications for Test Validity. *Educational and Psychological Measurement*, *32*(4), 1039–1044. https://doi.org/10.1177/001316447203200420
Jarrett, R. F. (1948). The Extra-Chance Nature of Changes in Students' Responses to Objective Test-Items. *The Journal of General Psychology*, *38*(2), 243–250. https://doi.org/10.1080/00221309.1948.9711785
Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling Answer Change Behavior: An Application of a Generalized Item Response Tree Model. *Journal of Educational and Behavioral Statistics*, *42*(4), 467–490. https://doi.org/10.3102/1076998616688015

Jeske, D., Backhaus, J., & Stamov Roßnagel, C. (2014). Self-regulation during e-learning: Using behavioural evidence from navigation log files. *Journal of Computer Assisted Learning*, *30*. https://doi.org/10.1111/jcal.12045

Juvina, I., & Oostendorp, H. van. (2006). Individual differences and behavioral metrics involved in modeling web navigation. *Universal Access in the Information Society*, *4*(3), 258–269. https://doi.org/10.1007/s10209-005-0007-7

Kim, Y. H., & Goetz, E. T. (1993). Strategic processing of test questions: The test marking responses of college students. In *Learning and Individual Differences* (Vol. 5, Issue 3, pp. 211–218). Elsevier Science. https://doi.org/10.1016/1041-6080(93)90003-B

Kinnebrew, J., Loretz, K. M., & Biswas, G. (2013). A Contextualized, Differential Sequence Mining Method to Derive Students' Learning Behavior Patterns. *EDM 2013*.

Kralisch, A., Eisend, M., & Berendt, B. (2005). *The Impact of Culture on Website Navigation Behaviour*.

LAK '11: Proceedings of the 1st International Conference on Learning Analytics and Knowledge. (2011).

Lawless, K. A., Mills, R., & Brown, S. W. (2002). Children's Hypertext Navigation Strategies. *Journal of Research on Technology in Education*, *34*(3), 274–284. https://doi.org/10.1080/15391523.2002.10782349

Lee, Y.-H., & Haberman, S. J. (2016). Investigating Test-Taking Behaviors Using Timing and Process Data. *International Journal of Testing*, *16*(3), 240–267. https://doi.org/10.1080/15305058.2015.1085385

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(1), 8. https://doi.org/10.1186/s40536-014-0008-1

Lynch, D. O., & Smith, B. C. (1972). *To Change or Not to Change Item Responses When Taking Tests: Empirical Evidence for Test Takers.*

Macgregor, S. K. (1999). Hypermedia Navigation Profiles: Cognitive Characteristics and Information Processing Strategies. *Journal of Educational Computing Research*, *20*(2), 189–206. https://doi.org/10.2190/1MEC-C0W6-111H-YQ6A

Makany, T., Engelbrecht, P., Meadmore, K., Dudley, R., Redhead, E., & Dror, I. (2007). *GIVING THE LEARNERS CONTROL OF NAVIGATION: COGNITIVE GAINS AND LOSSES*.

Marchionini, G., & Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. *Computer*, *21*(1), 70–80. https://doi.org/10.1109/2.222119

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*.

McClain, L. (1983). Behavior during Examinations: A Comparison of "A," "C," and "F" Students. *Teaching of Psychology*, *10*(2), 69–71. https://doi.org/10.1207/s15328023top1002_2

McEneaney, J. E. (2001). Graphic and numerical methods to assess navigation in hypertext. *International Journal of Human Computer Studies*, *55*(5), 761–786. https://doi.org/10.1006/ijhc.2001.0505

Mcnulty, J., Sonntag, B., & Sinacore, J. (2007). Test-taking Behaviors on a Multiple-Choice Exam are Associated with Performance on the Exam and with Learning Style. *Med Sci Educ*, *17*.

Mehdi Khosrow-Pour, D. B. A. (2010). Encyclopedia of Information Science and Technology, Second Edition. In M. Khosrow-Pour, D.B.A. (Ed.), *Encyclopedia of Information Science and Technology, Second Edition* (Second Edi). IGI Global. https://doi.org/10.4018/978-1-60566-026-4

Mostow, J., & Beck, J. (2006). Some Useful Tactics to Modify, Map and Mine Data from Intelligent Tutors. *Nat. Lang. Eng.*, *12*(2), 195–208. https://doi.org/10.1017/S1351324906004153

Mühlenbrock, M. (2005). *Automatic Action Analysis in an Interactive Learning Environment*.

Nagaoka, K., & Ueno, M. (2004). *Learning Log Database and Data Mining system for e-Learning OnLine Statistical Outlier Detection of irregular learning processes*.

Niederhauser, D. S., Reynolds, R. E., Salmen, D. J., & Skolmoski, P. (2000). The Influence of Cognitive Load on Learning from Hypertext. *Journal of Educational Computing Research*, *23*(3), 237–255. https://doi.org/10.2190/81BG-RPDJ-9FA0-Q7PA

OECD. (2013). *PISA 2012 Results: What Students Know and Can Do (Volume I)*. https://doi.org/https://doi.org/https://doi.org/10.1787/9789264201118-en

Ogata, H., Oi, M., Mouri, K., Okubo, F., Shimada, A., Yamada, M., Wang, J., & Hirokawa, S. (2017). Learning Analytics for E-Book-Based Educational Big Data in Higher Education. In *Smart Sensors at the IoT Frontier* (pp. 327–350). https://doi.org/10.1007/978-3-319-55345-0_13

Papanikolaou, K., & Grigoriadou, M. (2004). *Accommodating learning style characteristics in Adaptive Educational Hypermedia Systems*.

Paraskeva, F., Bouta, H., & Papagianni, A. (2008). Individual characteristics and computer self-efficacy in secondary education teachers to integrate technology in educational practice. *Computers & Education*, *50*(3), 1084–1091. https://doi.org/https://doi.org/10.1016/j.compedu.2006.10.006

Pechenizkiy, M., Trcka, N., Vasilyeva, E., Aalst, W. V, & Bra, P. D. (2009). Process Mining Online Assessment Data. *EDM*.

Rauterberg, M. (1999). *How to Measure Cognitive Complexity in Human-Computer Interaction*.

Richter, S. (2012). Learning Tasks. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1975–1979). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_342

Riding, R., & Cheema, I. (1991). Cognitive styles: An overview and integration. In *Educational Psychology* (Vol. 11, Issues 3–4, pp. 193–215). Taylor & Francis. https://doi.org/10.1080/0144341910110301

Riding, R., & Rayner, S. (2013). *Cognitive Styles and Learning Strategies* (1st ed.). David Fulton Publishers. https://doi.org/10.4324/9781315068015

Rindler, S. E. (1980). The Effects of Skipping Over More Difficult Items On Time-Limited Tests: Implications for Test Validity. *Educational and Psychological Measurement*, *40*(4), 989–998. https://doi.org/10.1177/001316448004000425

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions On*, *40*, 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/https://doi.org/10.1002/widm.1075

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

Salles, F., Dos Santos, R., & Keskpaik, S. (2020). When didactics meet data science: process data analysis in large-scale mathematics assessment in France. *Large-Scale Assessments in Education*, *8*(1), 7. https://doi.org/10.1186/s40536-020-00085-y

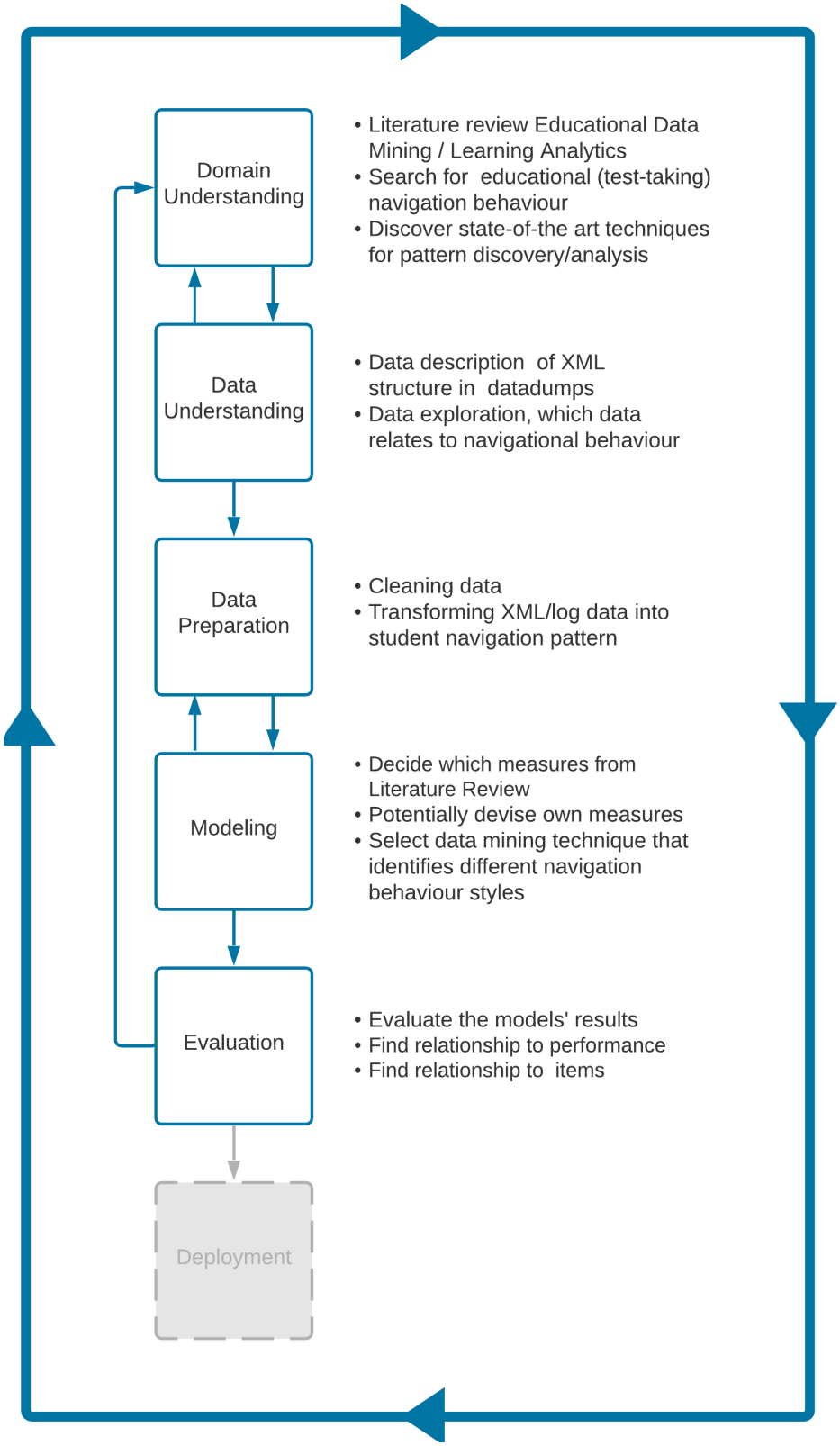
Scheuer, O., & Mclaren, B. (2012). Educational data mining. *Encyclopedia of the*

Sciences of Learning, 1075–1079.
Schnipke, D. L. (1995). *Assessing Speededness in Computer-Based Tests Using Item Response Times*.
Schnipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times With a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/https://doi.org/10.1111/j.1745-3984.1997.tb00516.x
Shu, Z., Hao, J., & von Davier, A. (2015). Analyzing Process Data from Game/ScenarioBased Tasks: An Edit Distance Approach. *Journal of Educational Data Mining*, *7*.
Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a Closer Look: An Exploratory Analysis of Successful and Unsuccessful Strategy Use in Complex Problems. *Frontiers in Psychology*, *10*, 777. https://doi.org/10.3389/fpsyg.2019.00777
Stenlund, T., Eklöf, H., & Lyrén, P.-E. (2017). Group differences in test-taking behaviour: an example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 4–20. https://doi.org/10.1080/0969594X.2016.1142935
Streeter, L. A., Vitello, D., & Wonsiewicz, S. A. (1985). How to tell people where to go: comparing navigational aids. *International Journal of Man-Machine Studies*, *22*(5), 549–562. https://doi.org/10.1016/S0020-7373(85)80017-1
Sukaviriya, P. N., & Foley, J. D. (1993). Supporting Adaptive Interfaces in a Knowledge-Based User Interface Environment. *Proceedings of the 1st International Conference on Intelligent User Interfaces*, 107–113. https://doi.org/10.1145/169891.169922
Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1
Tauscher, L., & Greenberg, S. (1997). How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, *47*(1), 97–137. https://doi.org/https://doi.org/10.1006/ijhc.1997.0125
Thuering, M., Hannemann, J., & Haake, J. (1995). Hypermedia and Cognition: Designing for Comprehension. *Commun. ACM*, *38*, 57–66. https://doi.org/10.1145/208344.208348
Tsianos, N., Lekkas, Z., Germanakos, P., Mourlas, C., & Samaras, G. (2008). *User-Centric Profiling on the Basis of Cognitive and Emotional Characteristics: An Empirical Study*. https://doi.org/10.1007/978-3-540-70987-9_24
van der Linden, W. J. (2008). Using Response Times for Item Selection in Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5–20. https://doi.org/10.3102/1076998607302626
van der Linden, W. J. (2009). Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, *46*(3), 247–272. https://doi.org/https://doi.org/10.1111/j.1745-3984.2009.00080.x
van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting Differential Speededness in Multistage Testing. *Journal of Educational Measurement*, *44*(2), 117–130. https://doi.org/https://doi.org/10.1111/j.1745-3984.2007.00030.x
van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265. https://doi.org/10.1007/BF02294800
Vasilyeva, E., De Bra, P., & Pechenizkiy, M. (2010). *Strict Order vs. Flexible Order Navigation in Online Testing*.
Vee, M.-H. N. C. (2006). *Understanding novice errors and error paths in Object-*

oriented programming through log analysis.
Vida, L., Maria, B., & Brinkhuis, M. J. S. (2021). *Speeding up without Loss of Accuracy: Item Position Effects on Examinees' Performance in University Exams*.
Von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, *2*, 9–36.
Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.
Wang, F.-H., & Shao, H.-M. (2004). Effective personalized recommendation based on time-framed navigation clustering and association mining. *Expert Systems with Applications*, *27*(3), 365–377. https://doi.org/https://doi.org/10.1016/j.eswa.2004.05.005
Wang, K. H., Wang, T. H., Wang, W. L., & Huang, S. C. (2006). Learning styles and formative assessment strategy: enhancing student achievement in Web-based learning. *Journal of Computer Assisted Learning*, *22*(3), 207–217. https://doi.org/https://doi.org/10.1111/j.1365-2729.2006.00166.x
Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2
Wise, S. L., Pastor, D., & Kong, X. (2009). Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice. *Applied Measurement in Education - APPL MEAS EDUC*, *22*, 185–205. https://doi.org/10.1080/08957340902754650
Witkin, H. A., Moore, C. A., Goodenough, D. R., & Cox, P. W. (1977). Field-Dependent and Field-Independent Cognitive Styles and Their Educational Implications. *Review of Educational Research*, *47*(1), 1–64. https://doi.org/10.2307/1169967
Wu, B. (2017). An Eye Tracking Study of High- and Low-Performing Students in Solving Interactive and Analytical Problems. *Educational Technology & Society*, *20*.
Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. *Applications of Latent Trait and Latent Class Models in the Social Sciences*, 89–98.

Appendices

Appendix A *Detailed version of this research adaptation of Chapman's CRISP-DM*



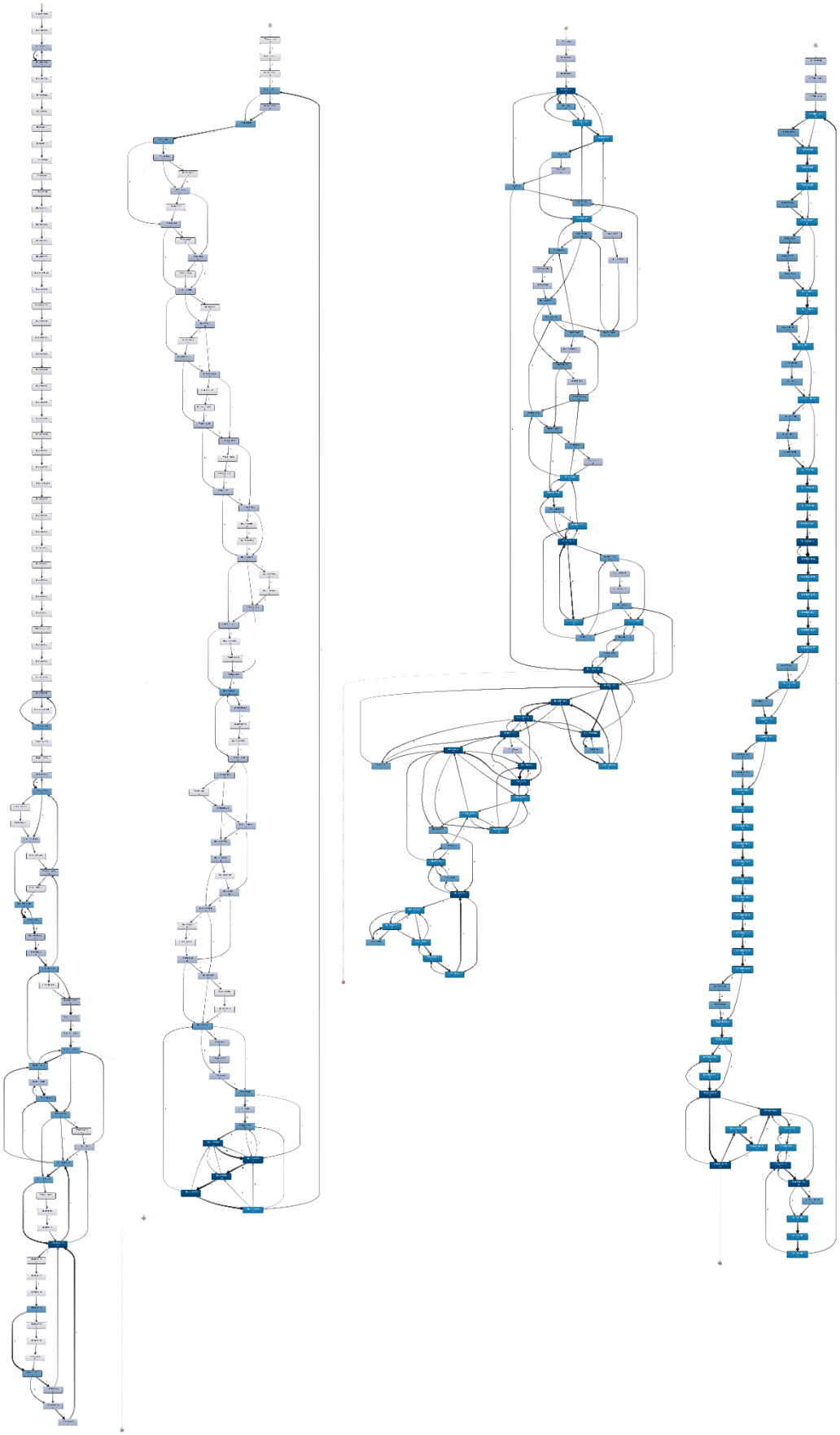
Appendix B *Overview EDM application categories*

(Baker & Yacef, 2009)	(Castro Espinoza et al., 2007)	(Scheuer & Mclaren, 2012)	(Bousbia & Belamri, 2014)	(Romero & Ventura, 2010)	(Bakhshinategh et al., 2018)
Improving student models	assessment of student's performance	Scientific inquiry and system evaluation	Student modelling	User/Student modelling	Predicting student performance
Improving domain models	course adaptation and learning recommendations based on the student's learning behaviour	Determining student model parameters	Domain structure analysis	Domain modelling	Detecting undesirable student behaviours
Studying the pedagogical support provided by the learning software	evaluation of learning material and educational web-based courses	Informing domain models	Generating recommendation	Recommending to students	Profiling and grouping students
Scientific research into learning and learners	feedback to both teacher and students in e-learning courses	Creating diagnostic model	Analyzing learner's behaviour	Scientific inquiry	Social network analysis
	detection of atypical students' learning behaviours	Creating reports and alerts for instructors, students and other stakeholders	Communicating to stakeholders	Providing feedback for supporting instructors	Providing reports
		Recommending resources and activities	Predicting students' performance and learning outcomes	Creating alerts for stakeholders	Creating alerts for stakeholders
			Maintaining and improving courses	Predicting student performance	Planning and scheduling
				Personalizing to students	Creating courseware
				Grouping/Profiling students	Developing concept maps
				Constructing courseware	Generating recommendation
				Planning and scheduling	Adaptive systems
				Parameter estimation	Evaluation
					Scientific inquiry

Appendix C *Description of variables found in cleaned data*

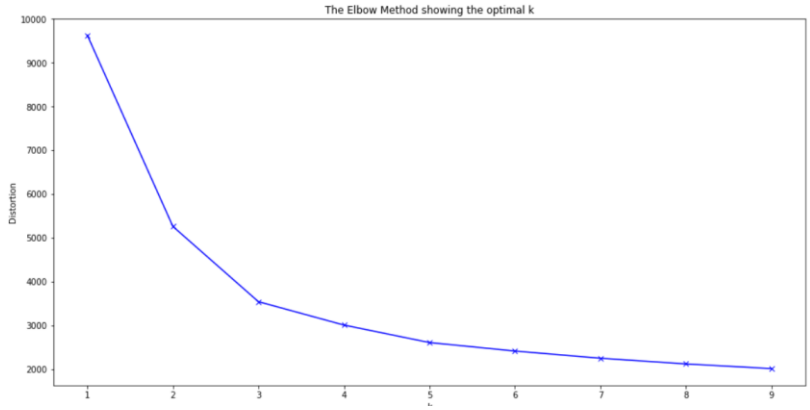
Variable	Description
Student	
Student_id	Identifier/number of student
StudentSex	Gender of Student
StudentDataofbirt	Student's date of birth
Test group	
Testgroup_id	Identifier of the test group. Will identify when and where a test session took place
TestgroupName	Name of the test group. Often the name of the school, but did differ
TestData	Date of test
Brin	Identifier of school
Test	
Test_id	Identifier of test
TestStatus	Status of the test. Whether or not the test was finished or abrupted
TestVersion_id	Identifier of the test version
Response	
ResponseIndex	The order in which the student took the test. Not to be confused with the item order presented by test.
Response	Answer to the question
ResponseTime	Amount of time in seconds to answer
Item	
Item_id	Identifier of question
Subject	The subject of the question (math, language, world orientation)
InfoItem	Boolean, if yes, this item contains instructions for the next set of questions. Does not require/output a response
ItemDifficulty	The predetermined difficulty level of the question
ItemType	Type of question (Multiple Choice)
ItemCardinality	The number of answers required (One or More)
Scoring Rule	
itemScore	Score of an answer to a question

Appendix D *Visualization of student's test-taking navigation behaviour*



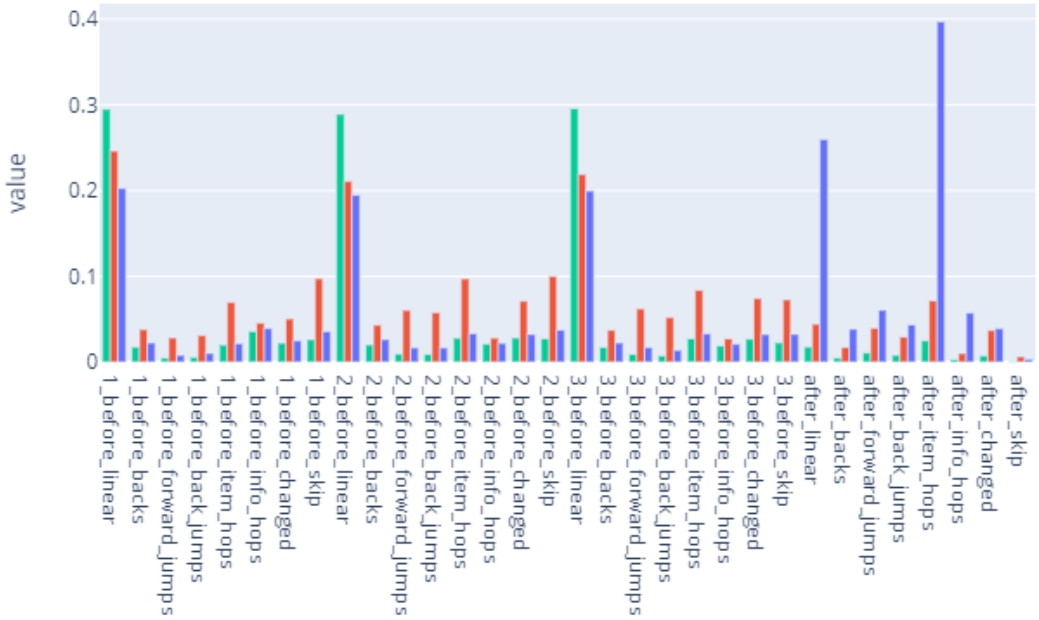
Appendix E Snippet of our k-means algorithm code in Python

```
from sklearn.preprocessing import MinMaxScaler
import plotly.graph_objects as go
import numpy as np
import plotly.express as px
import nbformat

scaler = MinMaxScaler(feature_range = (0,1))
scaler.fit(afname)
afname = scaler.fit_transform(afname)

inertia = []
K = range(1,10)
for k in K:
    kmeans = KMeans(
        n_clusters=k, init="k-means++",
        n_init=10,
        tol=1e-04, random_state=42
    )
    kmeans.fit(afname)
    inertia.append(kmeans.inertia_)

plt.plot(K, inertia, 'bx-')
plt.afnamelabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```




```
kmeans = KMeans(
        n_clusters=3, init="k-means++", n_init=10,
        tol=1e-04, random_state=42
    )
kmeans.fit(afname)
clusters=pd.DataFrame(afname,columns=afname_measures[['before_forward','before_backs','before_forward_jumps', 'before_back_jumps', 'before_item_hops', 'before_info_hops', 'before_changed', 'before_skip', 'after_forward','after_backs','after_forward_jumps', 'after_back_jumps', 'after_item_hops', 'after_info_hops', 'after_changed', 'after_skip']].columns)
clusters['label']=kmeans.labels_
afname_measures['label']=kmeans.labels_

bar=afname_measures[['before_forward','before_backs','before_forward_jumps', 'before_back_jumps', 'before_item_hops', 'before_info_hops', 'before_changed', 'before_skip', 'after_forward','after_backs','after_forward_jumps', 'after_back_jumps', 'after_item_hops', 'after_info_hops', 'after_changed', 'after_skip', 'label']].groupby("label").mean().reset_index()
bar["label"] = bar["label"].astype("category")
bar=pd.melt(bar,id_vars=["label"])
fig4 = px.bar(bar, x='variable', y='value', color='label', color_discrete_map={0:'#636EFA', 2:'#EF553B', 1:'#00CC96'}, barmode='group')
fig4.update_layout(showlegend=False)
fig4.show()
```

Appendix F *Groups of students clustered using all navigation measures, divided into 4 sections based on item position (blue = mostly linear, red = more activity before last item, green = more activity after last item)*



Appendix G *Count per groups of clustered tests per student ([0,0,0] = student showed behaviour 0 (primarily linear navigation) on all 3 tests)*

Group	count
[0, 0, 0, 0]	1324
[0, 0, 0, 1]	371
[0, 0, 0, 2]	132
[0, 0, 0]	7926
[0, 0, 1, 1]	230
[0, 0, 1, 2]	65
[0, 0, 1]	2801
[0, 0, 2, 2]	49
[0, 0, 2]	1178
[0, 0]	347
[0, 1, 1, 1]	156
[0, 1, 1, 2]	32
[0, 1, 1]	1798
[0, 1, 2, 0]	15
[0, 1, 2, 1]	5
[0, 1, 2, 2]	38
[0, 1, 2]	663
[0, 1]	127
[0, 2, 1, 1]	7
[0, 2, 2, 2]	42
[0, 2, 2]	634
[0, 2]	72
[0]	363
[1, 1, 1, 1]	69
[1, 1, 1, 2]	23
[1, 1, 1]	1202
[1, 1, 2, 2]	10
[1, 1, 2]	451
[1, 1]	124
[1, 2, 2, 2]	23
[1, 2, 2]	363
[1, 2]	57
[1]	121
[2, 2, 2, 2]	39
[2, 2, 2]	544
[2, 2]	74
[2]	59

Appendix H *Types of Answer Changes Made to Items of different difficulties in our dataset per clusters*

(n=245.113 changes)	Item difficulty			
Cluster 0	*Low*	*Moderate*	*High*	Total
Right to Wrong	0.7%	0.9%	1.1%	2.8%
Wrong to Right	22.0%	20.3%	17.5%	59.7%
Wrong to Wrong	8.3%	12.2%	17.0%	37.5%
Total	31.0%	33.4%	35.6%	100%

(n=293.848 changes)	Item difficulty			
Cluster 1	*Low*	*Moderate*	*High*	Total
Right to Wrong	1.1%	1.3%	1.5%	3.9%
Wrong to Right	21.1%	20.6%	17.7%	59.4%
Wrong to Wrong	7.7%	11.9%	17.1%	36.7%
Total	29.9%	33.8%	36.30%	100%

(n=89.256 changes)	Item difficulty			
Cluster 2	*Low*	*Moderate*	*High*	Total
Right to Wrong	1.5%	1.7%	2.1%	5.4%
Wrong to Right	20.2%	20.0%	19.8%	60.0%
Wrong to Wrong	7.0%	10.7%	16.9%	34.6%
Total	28.7%	32.5%	38.8%	100%