Clustering Sequential Navigation Patterns in Multiple-Source Reading Tasks with Dynamic Time Warping Method

Qiwei He[1*], Francesca Borgonovi[2,3], Javier Suárez-Álvarez [4]

[1] Educational Testing Service

[2] University College London

[3] Organization for Economic Co-operation and Development

[4] University of Massachusetts Amherst

*Author Note

Correspondence concerning this article should be sent to Qiwei He, Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA, email: qhe@ets.org

Abstract

Background: Data-driven investigations of how students transit pages in digital reading tasks and how much time they spend on each transition allow mapping sequences of navigation behaviors into students' navigation reading strategies.

Objectives: The purpose of this study is threefold: (1) to identify students' navigation patterns in multiple-source reading tasks using a sequence clustering approach; (2) to examine how students' navigation patterns are associated with their reading performance and socio-demographic characteristics; (3) to showcase how the navigation sequences could be clustered on the similarity measure by dynamic time warping (DTW) methods.

Methods: This study draws on process data from a sample of 16,957 students from 69 countries participating in the PISA 2018 study to identify how students navigate through a multiple-source reading item. Students' navigation sequences were characterized by two indicators: the page sequence that tracks the page transition path and the time sequence that records the time duration on each visited page. K-medoid partitioning clustering analyses were conducted on pairwise distance similarity measures computed by the DTW method.

Results and Conclusions: Students' navigation patterns were found moderately associated with their reading proficiency levels. Students who visited all the pages and spent more time reading without rush transitions obtained the highest reading scores. Girls were more likely to achieve higher scores than boys when longer navigation sequences were used with shorter reading time on transited pages. Students who navigated only limited pages and spent shorter reading time were averagely at the lowest rank of socio-economic status.

Implications: This study provides evidence for the exploration of students' navigation patterns and the examination of associations between navigation patterns and reading scores with the use of process data.

Key words: navigation sequences, dynamic time warping, sequence clustering, multiple-source reading tasks, PISA

Clustering Sequential Navigation Patterns in Multiple-Source Reading Tasks with Dynamic Time Warping Method

1. Introduction

Computer-based assessments enable the use of new sources of information to examine in detail not only how well test takers perform in the tasks they are presented with but also the strategies they employ to reach a solution. The use of these new sources of information in computer-based assessments is particularly valuable when assessing competencies that require interactive tasks and the use of technology. When reading proficiency is assessed, for example, students display varying degrees of mastery in the use of information and communication technologies (ICT) to access texts through search engines, use links and tabs, process information from multiple sources, evaluate the quality of information sources, detect possible conflicts, and resolve them. Process data, more specifically the sequences of actions performed by students when responding to the test, allow the identification of the navigation strategies used by students when interacting with texts in a digital environment. Most importantly, these data provide additional information beyond response data: they can help define how test takers solve a task and not just if they solve the task correctly or incorrectly (He, Borgonovi, & Paccagnella, 2019, 2021; He & von Davier, 2015, 2016). Data-driven investigations can aid the identification of navigation strategies by identifying meaningful sequences of navigation behaviors based on how students transit pages in digital reading tasks and how much time they spend on each transition. Sequences of navigation behaviors (the observed sequences of what test takers do when they are exposed to digital reading tasks) can be used to characterize navigation reading strategies (i.e. theoretical sequences of actions that define coherent plans to solve the task).

In today's global information society, extracting information through social media to solve a real-world problem or acquire knowledge online is key to work performance and successful participation in society (Gao et al., 2022). The capacity to navigate through digital platforms is increasingly recognized as a key reading skill. Navigation skills allow readers to access multiple-sources, "construct" their text, and retrieve relevant information from such texts (OECD, 2021a). Previous studies indicate that proficient readers of digital texts minimize the number of visits to irrelevant pages, locate relevant pages efficiently, and allocate time to the pages that contain the most relevant content (e.g., Lawless & Kulikowich, 1996; Salmerón & García, 2011). Analyses of how individuals navigate pages in digital texts combined with information on the accuracy with which they are able to extract meaning from accessing such pages could help identify how best to promote skill development in technology rich environments (Kinnebrew et al., 2013). Integrating the amount of time test takers spend on the pages they visit with the length and breadth of navigation patterns could also help researchers identify differences in search efficiency (e.g., Goldhammer et al., 2013; Vaughan & Dillon, 2006).

This paper uses process data containing information on the navigation behaviors and time use of 15-year-year-old students as they complete digital reading tasks to identify groups of students with similar navigation profiles. The aim of the analyses is to develop a deeper understanding of students' reading comprehension and to identify, by mapping students' navigation strategies through the traces left by their interaction with the testing platform, the cognitive processes that they use to solve digital text comprehension tasks. We do so by developing a novel approach to cluster dynamic action sequences and to measure similarities between different sequences and apply this approach to data from a reading task within the

multiple-source digital environment in the Programme for International Student Assessment (PISA) 2018 cycle. We conclude our analyses by examining differences in navigation profiles by gender and socio-economic groups.

1.1 Multiple source and dynamic texts in PISA reading tasks

Meaningfully and critically integrating information from multiple sources of information is vital for twenty-first-century literacy (Britt et al., 2018; List & Alexander, 2018; Magliano et al., 2017; Salmerón et al., 2018). For example, a reader interested in learning more basic scientific and health-related facts during the Covid-19 pandemic, such as whether using a face mask is effective in reducing the contagion rate, or the positive and negative consequences of vaccines and boosters, needs to be able to navigate through ambiguity and triangulate and validate viewpoints to make an informed decision about whether to use a face mask or get vaccinated. In an information-rich digital world, these types of texts, known as multiple sources and dynamic texts, are the norm and often present conflicting information. Despite the proliferation of research on strategies for comprehending and processing multiple source and dynamic texts (Cho et al., 2018; Ritcher & Maier, 2017; Van Meter et al., 2020), gaps remain and it is thus necessary to expand empirical evidence on how students navigate and integrate information from multiple sources and dynamic texts using computer-based assessment data across different types of texts, age groups, situations, and countries (Barzilai et al., 2018; Naumann, 2019; Naumann, 2015; Naumann & Salmeron, 2018).

International Large-scale Assessments (ILSA's) have often been at the forefront of innovations in test design and the use of analytic methodologies. PISA, for example, was among the first ILSAs to transition its operations to computer-based delivery among student populations

and to introduce computerized adaptive testing to assess the reading proficiency of 15-year-old students. The interactive nature of computerized assessments such as PISA makes them ideal candidates for analyses based on process data (Goldhammer et al., 2016; Vörös et al., 2021).

Compared to previous cycles, PISA 2018 put a greater emphasis on using multiple-source texts, i.e., texts composed of several units of text, created separately by different authors. While the availability of multiple sources of text does not necessarily imply greater difficulty, including multiple-source units helped to expand the range of higher-level reading processes and strategies measured in PISA. These include: searching for information across multiple documents, integrating across texts to generate inferences, assessing the credibility of sources, and handling conflicting information (OECD, 2021a). Unlike single-source texts that may have one definite author (or a group of authors), time of writing or publication date, and reference file or number, multiple-source texts are defined as texts with links, different authors, that have been published at different times, or bear different titles or reference numbers (OECD, 2019a). For example, the publicly released scenario-based reading unit, Rapa Nui (CR551) that was administered in PISA 2018 consists of three texts: a web page from a professor's blog, a book review, and a news article from a science magazine. In these multiple-text reading situations, readers must make decisions about which of the available text fragments is the most important, relevant, accurate, and truthful.

The PISA 2018 analytical framework also distinguishes between static texts – i.e. texts that have a simple organization and low density of navigational tools (e.g., scroll bars and tabs) – from dynamic texts – i.e. texts that feature a more complex, non-linear organization and a higher density of navigational devices (e.g., table of contents, hyperlinks to switch between segments of text or interactive tools such as in social networks). Dynamic texts give the reader a degree of

decision-making power over the construction of a narrative and the timing and sequence of interactions with the testing material that are not possible in static texts. Students have the flexibility to construct their own pathways and decide which information is important to be able to solve a task, and switch between pages. Therefore, analyzing students' navigation behaviors when students are presented with multiple sources and dynamic tasks that require constant interaction with the testing platform, enables analysts to map their problem-solving process.

1.2 Indicators of navigation behaviors in reading tasks

Previous studies have used the absolute or relative number of interactions, the length or breadth of navigation, and the total time spent on the task to measure search efficiency and observe differences across multiple experimental conditions (e.g., Mislevy et al., 2012; Baker & Yacef, 2009). For instance, Vaughan and Dillon (2006), used the time on task, the total number of page visits, and the number of category nodes visited at least once to identify how undergraduate students become more efficient in navigating information online after engaging in multiple sessions. Gao et al. (2022) used navigation behaviors based on how many times each web page was browsed by participants on a reading task to divide them into different problem-solving behavioral groups using data from the Programme for the International Assessment of Adult Competencies (PIAAC).

Solving tasks in complex interactive environments often requires the extensive use of navigation behaviors, and indicators of navigation behaviors are therefore useful to represent higher-level reading processes and strategies. For instance, in analyses of how proficiently test takers performed a web search task, Herder (2005) included a series of indicators of navigation behaviors including: perceived disorientation, page re-visitation ratio, percentage of back button

clicks among the navigational actions, the average number of times that a page was revisited (return rate), the average length of a path between connected pages (average connected distance), and median view time of pages. Similarly, analyses of the optional digital reading assessment that was administered in PISA 2009 (OECD, 2011; Naumann, 2015), three measures were developed to characterize test-takers' navigational behaviors: (1) number of page visits and revisits; (2) number of visits and revisits to task-relevant pages; and (3) number of relevant pages visited.

Analyzing complex navigation behaviors, such as detecting sequential patterns or computing association rules, is crucial for identifying information foraging patterns (Eichmann et al., 2020; Gabadinho et al., 2011; Liu et al., 2004). Navigation behaviors contained in clickstream data can be highly informative to identify navigation patterns and their effectiveness. (Eichmann et al., 2020, Hahnel et al., 2018; Jenkins et al., 2003). Recent studies have applied full-path sequence analysis to identify typical behavior patterns that best characterize the set of sequences of navigation pattern groups (Gao et al., 2022). Though the sequence distance similarity indicators hold the promise of describing navigation behaviors in a more comprehensive way, these indicators are not commonly used, mainly because of computational challenges (OECD, 2021a). In this study, we introduce a novel approach to calculate the navigation sequence and the sequence of time that was spent on each visited page using the dynamic time warping (DTW) method.

1.3 Evidence on socio-economic and gender disparities in reading

Many countries have experienced increasing income and wealth inequality in the past decades (OECD, 2015a; Saez & Zucman, 2016). Such increases in inequality occurred at a time

when in many countries educational opportunities were greatly expanded although, despite continued attention among researchers and policy makers, the reading and mathematics achievement of students continue to reflect their socio-economic background (see, for example, Coleman et al., 1966, Sirin, 2005). An extensive literature documents the institutional frameworks that are associated with wider or smaller achievement disparities and the social processes that shape existing disparities in achievement outcomes (Van de Werfhorst & Mijs, 2010). However, despite significant investments and political commitments in ensuring that schooling reduces socio-economic disparities in key foundation skills including reading large inequalities remain when the achievement of students with a different socio-economic status is compared (Borgonovi & Pokropek, 2021a; Pokropek, Borgonovi, & Jakubowski, 2015; OECD, 2019b).

Perhaps surprisingly, little is known about students with a different socioeconomic background approach solving reading items since such understanding could help inform educational interventions aimed at reducing socio-economic disparities. Socio-economically disadvantaged students may approach interactive reading tasks differently from their more advantaged peers and adopt less efficient navigation strategies. It has in fact been observed that socio-economically disadvantaged students have somewhat less access to digital devices and use such devices differently compared to their more advantaged peers (Borgonovi & Pokropek, 2021b; Notten, et al., 2009) and that school principals working in socio-economically disadvantaged contexts lament lacking information and communication technology (ICT) infrastructure (OECD, 2019b). Disparities in access to technology in school settings, coupled with disparities in access to effective teachers (OECD, 2018) could determine disparities in how well students with a different background develop emerging key reading skills such as effective

navigation. It is also important not to forget that social and emotional skills like academic self-concept have the potential to compensate for the effects of socioeconomic disparities on academic performance (Suarez-Alvarez et al., 2014). For example, in PISA 2018, disadvantaged students perceived the reading assessment as more difficult than advantaged students even after accounting for students' reading scores in 70 countries and economies (OECD, 2021a).

A second dimension over which inequalities in navigation strategies may unfold is gender. The literature indicates that while at school, girls generally outperform boys in reading (Buchmann et al., 2008; Cole, 1997; DiPrete & Buchmann, 2013; OECD, 2015b) and that they obtain higher grades than boys when they are assessed in language and writing skills (Voyer & Voyer, 2014). However, the size of the gender gap in reading differs depending on the age at which boys and girls are tested (Borgonovi, Choi & Paccagnella, 2021) and the type of test that is being used (Solheim & Lundetræ, 2018). In particular, the literature suggests that boys and girls have relative strengths and weaknesses when they are assessed under different administration conditions and when different types of assessment stimuli are used (Willingham & Cole, 2013; Borgonovi, 2022). For example, previous studies indicate that the gender gap in reading in favor of girls is larger for open-ended questions (Beller & Gafni, 2000), and differs depending on cognitive processes and aspects of reading and other types of text characteristics (Solheim & Lundetræ, 2018). Furthermore, analyses suggest that the gender gap in reading depends on how long reading tests are since girls tend to have more stable levels of accuracy in long and cognitively demanding assessments than boys (Borgonovi & Biecek, 2016).

Of relevance to our work is the finding that the gender gap in reading measured in international large-scale assessments varies depending on whether the assessment considers print reading or digital reading with the gender gap in favor of girls being larger for print reading

(Borgonovi, 2016). It has been speculated that such differences might arise because digital reading requires greater visual-spatial ability (Lee, 2007) and because girls may have poorer navigation skills and may be less interested in digital texts than boys (Zhou, 2014). These differences, however, can also be associated with the opportunity to learn effective reading strategies in school. PISA data show almost two-thirds of the association between gender and reading performance can be accounted for by the difference between boys' and girls' knowledge of effective reading strategies (OECD, 2021a). Our analyses contribute to this growing body of literature examining what factors shape the gender gap in reading proficiency and, especially in aspects of reading proficiency that are becoming increasingly relevant for labor market participation and social integration.

1.4 Sequence-based features and distance similarity measures

Sequence-based features in process data analysis are primarily grouped into two categories: mini-sequences that are disassembled from a long sequence and measures of similarity computed by distances of pairs of full sequences. The mini-sequences are usually represented by n-grams, that is, a contiguous sequence of n items from a given sequence such as clickstream, text, or speech (He & von Davier, 2016). Features derived from clickstreams comprise: a) generic features commonly used in sequence mining or natural language processing (e.g., n-grams as in He & von Davier, 2015, 2016; Ulitzsch et al., 2022); b) task-specific features, created based on subject-matter knowledge on behavioral patterns to be expected on the task (Chen et al., 2019; Salles et al., 2020); or c) a combination of the two (Han et al., 2019; Liao et al., 2019; Qiao & Jiao, 2018). These features are then fed to classifiers or prediction models or analyzed using sequence mining techniques to identify features that best distinguish correct from

incorrect clickstreams (Ulitzsch et al., 2022).

Sequence distance functions are designed to measure sequence (dis)similarities in sequence mining. A common approach for detecting patterns in action sequences consists of converting the information contained in action sequences into distance measures (Dong & Pei, 2007). In the context of problem-solving processes, sequence distance measures can be defined to describe how action sequences differ either from each other (Tang et al., 2020; Ulitzsch et al., 2021; Gao et al., 2022) or with respect to pre-defined sequences (Hao et al., 2015; He, Borgonovi, & Paccagnella, 2019, 2021).

Character alignment-based distance functions are broadly used in sequence proximity metrics. These algorithms can be local window-based or whole sequence based; they can also be edit distances or more general pairwise similarity score-based distance (Tang et al., 2020; He et al., 2021; Dong & Pei, 2007). For instance, the edit distance function, also called the Levenshtein distance (Levenshtein, 1965, 1966), between two sequences $S1$ and $S2$ defines the minimum number of edit operations (i.e., deletion, insertion, and substitution) that are needed to transform $S1$ into $S2$ (Jurafsky & Martin, 2009). Hao et al. (2016) applied the edit distance method to compute the similarity between action sequences to identify the typical strategies by correct and incorrect groups to solve a scenario-based digital task. The Hamming distance between two sequences is limited to cases where the two sequences have identical lengths and is defined as the number of positions where the two sequences are different (Hamming, 1950). The Longest Common Subsequence algorithm (LCS; e.g., Hirschberg, 1975, 1977) identifies the longest subsequence that is common to two strings. The length of the LCS is defined as the degree of closeness between the two strings. Sukkarieh et al. (2012) used the LCS to cluster students' response sequences with high similarity and provide automatic scoring in multiple language

environment. He et al. (2021) employed the LCS to compute the similarity between the predefined sequence and individual observed sequence to explore how far the respondent's solution was away from the optimal path (i.e., the action sequence of correct solution with the minimum number of actions). Ulitzsch et al. (2021) extended the LCS application by integrating the time intervals between actions to better understand respondents' unusual test-taking behaviors - such as long-time pause and speedy skipping - to assist in pinpointing the potential reason for their success and failure in a digital task. Gao et al. (2022) used neighborhood density (Gabadinho et al., 2011) as a representativeness criterion for sorting candidate representative sequences and chose the optimal matching between sequences of spells (OMspell) algorithm (Studer & Ritschard, 2014) to measure similarities between sequences of navigation pages, thus resulted in four homogenous groups of navigation patterns.

In this study, we used the DTW method (Sakoe & Chiba, 1978), an alternative algorithm to compute the similarity between time-series sequences via dynamic programming. It was first developed by the speech recognition community to handle the matching of non-linearly expanded or contracted signals. The algorithm features in finding the optimal path through a matrix of points representing possible time alignments between the signals. In the context of navigation pattern exploration, each visited page could be recorded in a sequence and the time spent on each visited page could also be recorded into a sequence. The similarity between pairs of page sequences and time sequences could be computed through the DTW algorithm (which will be explained in detail in the Methods section).

The information contained in distance measures developed with these methods can be further aggregated by employing exploratory dimensionality reduction techniques such as principal component analysis, hierarchical clustering (Hao et al., 2015), and multidimensional

scaling (Tang et al., 2020).

1.5 The present study

The purpose of this study is threefold: first, to identify students' navigation patterns in multiple-source reading tasks using a data-driven approach (i.e., using their sequence to complete the task); second, to showcase how navigation sequences could be clustered on DTW similarity measure to identify behavioral navigation patterns, and third, to examine how students' navigation patterns are associated with their reading performance and their socio-demographic characteristics (i.e., gender and socio-economic status). Specifically, this study pursues to answer three research questions:

1. What are the representative navigation patterns (i.e., page transition and time spent on page transition) when students solve a multiple-source required reading task?
2. What is the association between navigation patterns and reading proficiency?
3. To what extent do students' navigation patterns differ systematically by gender and socio-economic status?

We answer these questions drawing on process data from students participating in PISA 2018 and we identify their navigation behaviors in one of the multiple-source required tasks (CR551Q11) in the computer-based reading unit "Rapa Nui".

To do so, we first developed two sequential indicators of navigation activity from process data: the navigation sequence that tracks the page transition path and the time sequence that records the time duration on each visited page. We then constructed two sequence similarity matrices by computing the distance of page sequence and time sequence respectively for each student pair with the dynamic time warping method. Next, to identify the optimal number of

clusters of sequences, we set the initial number of clusters from $n = 2$ to $n = 10$, therefore a series clustering analyses were conducted. We identified four representative navigation transition patterns and four time-allocation patterns on transition pages and examined the association with students' reading skills and socio-demographics variables (i.e., gender and socio-economic status).

2. Methods

2.1 Sample

We used a subsample of 16,957 students from 69 countries participating in PISA 2018 who executed at least one navigation activity (i.e., visited at least one page beyond the default homepage) in one example task (CR551Q11) in the reading Rapa Nui unit. The reason to exclude students who did not transit to any other pages beyond homepage was that their navigation length was null. The null value in navigation length could not derive a similarity measure against other sequences, thus had to be discarded. Compared with the full sample that was assigned to the Rapa Nui reading unit[1] in PISA 2018, 66.3% of the students[2] showed no navigation activities in the task CR551Q11 (OECD, 2021a). As in this study, we focused on the remaining approximately one-third of students who had at least one navigation activity, the sample reflects students who are active in navigation and it is therefore not representative of the general population of 15-year-old students that participated in PISA. In particular, as indicated in the 21st Century Readers report (OECD, 2021a), students who did not engage in any navigation

[1] The item CR551Q11 is the last item in the reading unit Rapa Nui. Given the time limits within the reading module, students had higher chances to skip the last items towards the end of the test. A higher rate of nonresponse (30.6%) and non-navigation (66.3%) was expected in this item, which might bring bias to the sample (favorable to the students with active navigation behaviors) compared with the full sample in PISA.

[2] The full sample of 76,270 students were the respondents from 70 countries (including Cyprus) assigned to the Rapa Nui unit including Cyprus. In this study, the sample of Cyprus was not used because of confidentiality concerns.

activity had lower reading achievement than other students. Consequently, the sample included in this study is biased towards students with higher levels of reading proficiency: the average reading score of the subsample of students in this study was 579 points, 62 points higher than that of the full sample of students who were administered the Rapa Nui reading unit. The sample is also different with respect to background characteristics such as gender and socio-economic status: 56.1% of students in our sample were girls compared to 52% of students that were presented the Rapa Nui unit. Similarly, students in our sample had a more advantaged socio-economic status, as indicated by the average value on the index of economic, social, and cultural status (ESCS), a composite index that reflects the parental occupation and educational attainment of participating students as well as the possessions students had in their households. The index takes a value of 0 across OECD countries and has a standard deviation of 1 with higher values reflecting a more advantaged background. The average ESCS index in our sample was 0.245 while across students presented with the reading unit this was -0.05.

2.2 Instruments

This study focuses on a scenario-based reading unit "Rapa Nui", which was developed with multiple-source text environments in PISA 2018 reading assessment. As the screenshots of one item (CR551Q11) in the reading unit Rapa Nui exhibited in Figure 1, this reading unit consists of three text sources: a webpage from the professor's blog, a book review, and a news article from an online science magazine (OECD, 2019a). When exploring the multiple source texts, readers must make decisions as to which of the available pieces of text is the most important, relevant, accurate or truthful (Rouet & Britt, 2011). The Rapa Nui unit consists of seven items ranging from moderate to high difficulty. The first five items (item 1 to item 5) are

items with single source texts. Students are instructed to complete the task using a single page as reference. Navigation to other pages that can be accessed by students is optional. The last two items (item 6 and item 7) require the use of multiple source texts. Each item instructs students to refer to all three sources to complete the task, requiring navigation to other pages beyond the default homepage.

****

[Insert Figure 1 around here.]

****

Given concerns that the length of navigation sequences needs to be sufficiently long (i.e., at least one navigation path) to identify reasonable navigation patterns, this analysis focused on the last item (CR551Q11) in the Rapa Nui unit, because, within the sample of students participating in PISA 2018, this was the item in which students made most page transitions (2.88 pages on average), and spent the longest reading time (159.5 seconds on average to explore all the text sources).

As shown in Figure 1, our target item is an open response item coded by human raters. This item ranks at a moderate-high difficulty level (Level 4) and focuses on evaluating students' capacity to detect information and handle conflicts. The information necessary to solve the item was located on multiple pages (blog, book review, and science news), meaning the transition among pages to identify the required information is crucial. To solve this item, students must integrate information from a range of theories presented in the texts and decide which theory to support, since such theories are at odds with one another. Successfully solving the item entails supporting any of the presented theory or neither as long as the answer acknowledged the need for additional research. (OECD, 2019a).

The reading unit Rapa Nui was included in the second unit of a testlet in the high difficulty module of the PISA and the item in this analysis was the last in the unit. As a result, the average nonresponse rate in the Rapa Nui unit was 16% and, in the item, CR551Q11 it was higher as 31%. Regarding students' performance in the target task, for the subsample used in this study, 65% of students got a full credit, 30.2% of students got no credit, and 4.8% of students were recorded as having missing responses (i.e., students visited at least one page beyond the default homepage, though didn't input a response).

2.3 Dynamic time warping method

The DTW method is one of the similarity distance measures which can be used to assess how similar two sequences are, especially when data entails a time-series format. It has been widely applied to problems in economic and sales forecasting (e.g., Arya et al., 2021; Chang et al., 2008), speech recognition (e.g., Permanasari et al., 2019; Amin & Mahmood, 2008), and music rhythm identification (e.g., Ren et al., 2016; Guo & Siegelmann, 2004). Unlike data types in traditional databases where the similarity of distance definition is straightforward, the distance between time series needs to be carefully defined in order to reflect the underlying proximity of these specific data, which is usually based on shapes and patterns (Kurbalija et al., 2011). For instance, in the stock market analysis, the stock curves by two companies may follow similar patterns but show peak occurrences at different time points. Following traditional approaches, for instance, using the Euclidean distance, to calculate the similarity measure between the two stock curves, the distance would yield very large distances between the two sequences, because it ignores that the shape of the two curves is very similar but located at a different pace. Analogously, the sequential process data resulting from clickstream and navigation sequences

echo the same needs. When we put the navigation sequence along a time axis, one student's navigation path could be similar to another, but with a different time pause at each action or page. Find the optimal warping path between the two sequences can help to reflect the appropriate similarity measures between the sequences.

Given two sequences $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_m\}$ with the same or different lengths, a warping path W is an alignment between X and Y, involving one-to-many mappings for each pair of elements. The cost of a warping path is calculated by the sum of the cost of each mapping pair. Furthermore, a warping path contains three constraints: (1) endpoint constraint: The alignment starts at pair $(1,1)$ and ends at pair (N, M); (2) monotonicity constraint: The order of elements in the path for both X and Y should be preserved in the same, original order of X and Y, respectively; (3) Step-size constraint: The difference of index for both X and Y between two adjacent pairs in the path needs to be no more than one step. In other words, pair (x_i, y_j) can be followed by three possible pairs including (x_{i+1}, y_j), (x_i, y_{j+1}) and (x_{i+1}, y_{j+1}).

DTW is a distance measure that searches the optimal warping path between two series. We first construct a cost matrix C, where each element $C(i,j)$ is a cost of the pair (x_i, y_j), specified by using a distance function. DTW is calculated based on dynamic programming. The initial step of DTW algorithm is defined as:

$$DTW(i,j) = \begin{cases} \infty & if\ (i = 0\ or\ j = 0)\ and\ i \neq j \\ 0 & if\ i = j = 0 \end{cases} \quad (1)$$

The recursive function of DTW is defined as

$$DTW(i,j) = min \begin{cases} DTW(i-1,j) + w_h C(i,j) \\ DTW(i,j-1) + w_v C(i,j) \\ DTW(i-1,j-1) + w_d C(i,j) \end{cases} \quad (2)$$

where (w_h, w_v, w_d) are weights for the horizontal, vertical and diagonal directions, respectively. $DTW(i,j)$ denotes the distance or cost between two sub-sequences $\{x_1, x_2, \dots, x_i\}$ and $\{y_1, y_2, \dots, y_j\}$, and $DTW(N,M)$ indicates the total cost of the optimal warping path.

For example, as shown in Figure 2, in the two sequences $A = \{1,2,3,4,5\}$, $B = \{1,2,3,4,3,2,5\}$, we first constructed a distance matrix. To calculate the value of each cell, we followed a combination of formula (1) and (2) as below:

$$dtw(i,\ j) = \left|A_i - B_j\right| + \min\left(D[i-1, j-1], D[i-1, j], D[i, j-1]\right) \quad (3)$$

****

[Insert Figure 2 around here.]

****

For example, to get the value in cell on the Column 2 Row 5 (in reverse order), that is highlighted by a red box in Figure 2, we calculated $dtw(5,\ 2) = |5-2| + \min(6,3,10) = 3 + 3 = 6$. After the whole matrix is developed, the shortest path (highlighted in yellow) starting from the diagonal corner was identified. We added up the shortest path to get the DTW distance similarity score between the two sequences as $3 + 3 + 1 + 0 + 0 + 0 + 0 = 7$.

It is noted that the DTW applies to numeric sequences, in which all the elements are numbers rather than categorical values in other sequence measures such as the longest common subsequence (He et al., 2021). Due to this situation, it is helpful to label each visited page or clickstream action as an ordered number to be applied in the DTW computation.

2.4 K-medoid partitioning clustering method

Based on the pairwise distance similarity matrix, we employed the k-medoid partitioning

clustering method (Kaufman & Rousseeuw, 1990) to cluster the sequences into homogenous groups. The partition clustering method aims to break the dataset into groups by minimizing the distance between points labeled to be in a cluster and a point designated as the center of that cluster (Jurafsky & Martin, 2009). In contrast to the k-means algorithm that are commonly used in clustering analysis (e.g., He, Liao, & Jiao, 2019), k-medoids chooses actual data points as centers (medoids or exemplars), and thereby allows for greater interpretability of the cluster centers than in k-means where the center of a cluster is not necessarily one of the input data points but the average between the points in the cluster. As the target in our study was to identify the typical navigation patterns, the k-medoid method is more favorable than k-means to interpret the homogeneous patterns. Furthermore, because k-medoids minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances, it is more robust to noise and outliers than k-means, which could help identify the aberrant navigation patterns more easily.

K-medoids is a classical partitioning technique of clustering that splits the data set of n objects into k clusters, where the number k of clusters assumed known a priori. In this study, we set k as 2 to 10 to as a priori and set the optimal number of clusters k with the silhouette index (Rousseeuw, 1987). The silhouette index ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. A score of 1 denotes the best meaning that the data point is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1. Values near 0 denote overlapping clusters. The centroid sequence derived from each cluster represents the sequence patterns of page transition and time allocation in the transition in each cluster.

2.5 Procedure

The present study was conducted in three steps. In the first step, we extracted two sequence indicators – the page-sequence that tracked the page transition path, $P_i = (p_1, p_2, \dots, p_n)$ and the time sequence that recorded the time duration on each visited page, $T_i = (t_1, t_2, \dots, t_n)$ from each student's navigation path on the item-level and recorded in a sequence manner. In the second step, we measured the similarity with DTW method between each pair of page sequences $DTW(P_{ij})$ and time sequences $DTW(T_{ij})$ respectively. The pairwise distance similarity of all students' page sequences and time sequences was written into two distance matrices **P** and **T** respectively. Based on the two matrices, we employed the k-medoid partitional clustering method to identify the optimal number of clusters of navigation patterns. In the third step, we further constructed a joint cluster membership by combining the page and time cluster results for each student. For instance, if there were four clusters derived from the page sequences and four clusters were derived from the time sequences respectively, the clusters with joint characteristics from both page and time sequences were constructed as $4 \times 4 = 16$ clusters. The students' reading competence was then associated with the cluster memberships of navigation patterns. The students' reading profile was further examined by background variables (i.e., gender and ESCS) by each cluster membership with a one-way analysis of variance statistics (ANOVA) with Bonferroni post-hoc correction.

3. Results

3.1 Navigation patterns

To answer the first research question, we executed clustering analysis on the sequence of page transitions and sequence of time spent on each transition, respectively. As shown in Figure 3, the silhouette index showed the maximum value, i.e., $S = 0.6219$ for page sequence

clustering and $S = 0.4235$ for time sequence clustering respectively, when the optimal number of clusters was set at 4 (i.e., $k = 4$). Therefore, we targeted at four typical navigation paths, and four time-sequences in dynamic navigation and visualized them with an extraction from the cluster centroids.

****

[Insert Figure 3 around here.]

****

Figure 4 exhibits four typical navigation paths that students employed when solving the multiple-source required reading task CR551Q11. The first page-cluster (P1) showed that students typically transited to only one page (i.e., book review) beyond the default homepage (i.e., blog) in the navigation path. The third cluster (P3) followed a similar pattern but the transit was to a different page (i.e., science news) beyond the default homepage (i.e., blog) in the navigation path. Though the content of visited pages was different in clusters P1 and P3, these two clusters both showed a limited navigation behavior. These two patterns were not combined into one group because, in the item interface, the tab "book review" was located closer to the "homepage" tab than the "science news" tab, thus influencing students' choice of navigation path. In contrast, cluster P2 involved students who employed a longer navigation path and reached to all the three pages to get information to complete the task. Typically, students in P2 transited to the science news page directly after reading the homepage (i.e., blog), and then transited to the book review page, followed by a revisited path to the science page. Cluster P4 showed similar pattern as P2, but with an even longer sequence because revisit behavior involved at least two pages in a row.

****

[Insert Figure 4 around here.]

****

Table 1 reports an overview of descriptive statistics by each navigational pattern cluster. As shown in Table 1, Cluster P1 has the biggest sample size (65.4%): a majority of students followed the most straightforward navigation path, i.e., transited only to one page, the book review page (the closer tab to the homepage). Although cluster P3 also showed the limited navigation path to only one page (i.e., science news page), only 20% of students employed this navigation pattern. Compared with students in Cluster P1, students in P3 might give a serious thought on which page should be explored to find the appropriate information to solve the task, rather than conveniently choosing an adjacent page to explore further. This interpretation might also explain why the P3 group had a higher average reading score (585) than P1 (574), although both of the two groups transited to only one page. A smaller number of students followed the navigation sequence pattern in P2 (11.5%) and P4 (2.6%), which revealed a longer navigation sequence embedded with revisit actions.

****

[Insert Table 1 around here.]

****

Figure 5 presents the typical sequence of time spent on visited pages, especially the time spent on the homepage and the first transited page[3]. The first time-cluster (T1) shows that students typically spent around 25 seconds on the homepage and showed a speedy transition to

[3] It is noted that the time clusters showed the reliable patterns from the homepage to the first transition page. Although the whole sequence of time allocation to different pages were input to the DTW algorithm, only the reliable patterns could be extracted and interpreted.

the next page but stayed only 2 seconds for a quick skimming before they put effective time on the third page to read for around 80 seconds. The second cluster (T2) showed students typically spent a significantly long time on the homepage (around 80 seconds) but used only 20 seconds in the transited page, a possible indicator of rush reading or quick skimming. The third cluster (T3) displayed the longest reading time in total but allocated a shorter time to the homepage (around 20 seconds) and transited to another page and read for a longer time for around 160 seconds. The last cluster (T4) showed the shortest reading time when completing the reading task. Typically, students in this group spent around 10 seconds on the homepage and 50 seconds in the transited page.

The sample size was relatively balanced by time sequence clusters. The biggest group, 34.8% of students, followed the T4 pattern, where students spent a short time reading the homepage and transited page. The second large group was T1: 28.1% of students. There were 23.5% of students clustered into the T3 pattern, and 13.6% of students in T2.

****

[Insert Figure 5 around here.]

****

3.2 The association between navigation patterns and reading proficiency

To answer the second research question, we examined the association between navigation patterns and students' overall proficiency in reading. ANOVA results indicate that the reading proficiency score was significantly different among the four page transition clusters ($F(3, 16953) = 63.267, p < 0.001$). The measure of effect size η^2, i.e., the percentage of the variance in the dependent variable (i.e., reading proficiency score) explained by the independent

variable (i.e., page transition sequence clusters), was 0.11. It suggests that 11% of the difference in reading proficiency score could be explained by the page navigation patterns. Table 1 shows, that in the context of page transition pattern, the students in P2 who adopted a multiple-page navigation approach with a focus on the scientific news and book review pages got the highest average reading score (599 points). In contrast, students in P1 who only navigated to the book review page obtained the lowest reading score (574 points). P3 and P4 did not show a significant difference in reading score in the Bonferroni test.

The reading proficiency score among the time clusters also shows a significant difference in ANOVA test $F(3, 16953) = 147.706, p < 0.001, \eta^2 = 0.25$. It suggests that 25% of the difference in reading proficiency score could be explained by the patterns of time spent on navigation pages. The students in T3 who allocated time primarily on the transited page had the highest average reading score (598 points) while students in T4 who typically spent a short time on the homepage and transited page had the lowest average reading score (564 points). The average reading score was not significantly different between T1 (i.e., using a speedy transition and skimming before putting significant time in reading the transited page) and T2 (i.e., allocated time primarily on the homepage).

We further examined the association between the joint navigation patterns in page- and time-cluster and reading proficiency scores. As shown in Figure 6, the cluster PT14 – that is, a combination of patterns P1 and T4, a short time on both initial page and transited page as well as only one transition to the book review page, resulted in the lowest reading score (557 points) across all 16 groups. In contrast, cluster PT43, that is, the pattern of a long sequence with multiple navigations across different pages and revisits pattern and spending a significant amount of time reading the transited page showed the highest reading performance score (612 points). A

substantial reading performance gap (55 points) was found between these two groups. This finding stresses the importance of examining students' strategies in the reading and problem-solving process.

****

[Insert Figure 6 around here.]

****

3.3 Disparities by gender and socioeconomic background in navigation patterns

To answer the third research question, we examined the relationships between navigation patterns and the sociodemographic characteristics of respondents (i.e., gender and socio-economic status). Columns 7 and 8 in Table 1 present the percentage of girls and boys employing the different navigation patterns. Results reveal few in preference of navigation page paths by gender, while the differences in their time allocation patterns were more prominent. Boys were four percentage points more likely than girls to be in the T4 pattern, an indication that boys were more likely to spend a shorter time reading the homepage and transited page to complete the reading task. By contrast, girls were 5 percentage points more likely than boys to be in the T3 group, indicating that girls were more likely to spend longer time reading on average, and, especially to spend their time reading the transited page.

We also examined the gender gap in the reading scores that were associated with each navigation pattern. As shown in Figure 7, in the joint navigation pattern clusters PT24, PT41, PT42 and PT44, girls scored by over 20 points higher than boys in the same pattern group. This suggests that even when boys and girls employed the same navigation patterns, they achieved at a different level. This might be related to the engagement status of boys and girls when adopting

the same navigation patterns. For instance, students grouped in clusters PT24 and PT44 engaged in long navigation sequences but short reading time and transition time. Girls might have a very efficient way to read text sources and transit across multiple pages. Conversely, boys' behavior might reflect quick reading and multiple transition, and the fact that they lose engagement by aimlessly quick switching across pages. Interestingly, boys were more likely to achieve in the highest range of proficiency score than girls, especially in the PT22 and PT43 patterns. These two patterns implied that students engaged in long navigation sequences and revisit to certain pages, and devoted significant time in reading either homepage (PT22) or transited page (PT43). This implies that boys who seriously navigated through multiple pages and spend significant time in reading the required pages scored, on average, between 5 and 10 points higher than girls who engaged in the same behavior. However, because the sample size in these two groups (PT22 and PT43) was relatively small in this study, caution needs to be taken to avoid drawing general conclusions.

****

[Insert Figure 7 around here.]

****

In order to examine the association between navigation patterns and socio-economic status, we mapped the ESCS index to each cluster (see the last three columns in Table 1) and ran the ANOVA analyses. Results revealed significant socio-economic differences in pattern P1 ($F(3, 16665) = 7.793, p < 0.001$). A mild effect size $\eta^2 = 0.07$ was found, suggesting the 7% of the variance in ESCS could be explained by the page navigation pattern. Students who followed the P1 pattern (i.e., transited only to the straightforward book review page) were in a lower range of the ESCS index and achieved the lowest reading score. Similarly, in the time

clusters, as shown in ANOVA analysis, $F(3, 16665) = 30.838$, $p < 0.001$, $\eta^2 = 0.01$, a significant difference was found among the four timing clusters, but the effect size was relatively small to explain the variance in the ESCS. The pattern T4, i.e., the shortest time in reading and navigation, also featured students with a significantly lower value on the ESCS index. Further, the average ESCS of students in PT14 (0.14) and PT44 (0.12) was the lowest out of all 16 combinations. Both PT14 and PT44 were involved short navigation and limited reading time, although the sequence of transition followed totally different patterns: PT14 took the shortest one transition which PT44 took the longest sequence. Students in the PT43 pattern, the pattern featuring the highest reading scores, was the pattern with the highest average ESCS index.

4. Discussion and Conclusion

As digital technologies and their use become pervasive, teenagers are increasingly required to apply their skills to read digital material to solve problems on computers and need to master how best to access and use information delivered on digital technologies to be successful in future jobs and society at large (OECD, 2021b). Education systems must respond to these changes by equipping students with the skillset that is needed to make the most of the information available on the internet. This study draws on sequential process data from one multiple-source reading task (CR551Q11) in the released unit 'Rapa Nui' in PISA 2018 to showcase how students' navigation strategies can be identified using a data-driven approach, namely sequence clustering and the DTW similarity measure. Education systems aim to develop skills among all students, ensure that all students achieve at least minimum standards of achievement, and reduce the impact of socio-economic and demographic factors on achievement. These results can also be helpful for teachers to understand what strategies students use in

solving literacy tasks and better support students' learning in reading, navigation, and information-gathering. In this study we analyzed whether students' navigation behaviors differed among girls and boys and among socio-economically advantaged and disadvantaged students and identified differences in accuracy in the test as a function of navigation behavior.

Our results indicate that navigation sequence patterns were moderately associated with students' reading proficiency levels: on average, students who visited all the pages and spent more time reading without rush transitions scored higher in reading than those students with less focused navigation. In contrast, students who limited their navigation to only one page beyond the initial default page and spent only a tiny amount of time on that page scored the lowest in reading. These results can also be helpful for teachers to understand what strategies students use in solving literacy tasks and better support students' learning in reading, navigation, and information-gathering. Crucially, our study revealed gender differences among students who employed the same navigation patterns: on average, girls were more likely to achieve higher reading scores than boys when longer navigation sequences with revisit patterns were used with shorter reading time on transited pages, while boys were more likely to achieve higher scores than girls when they spent longer time reading either homepage or transited page along with comprehensive navigation paths through the multiple pages. We also found that socio-economically disadvantaged students were more likely to use limited navigation activities and spend shorter reading time in both homepage and transited page than their more advantaged peers.

But what do these findings mean for schools and policymakers? On the one hand, the specialized literature seems to agree that reading from paper leads to better reading comprehension than digital reading, especially when there is a time constraint (Clinton, 2019;

Delgado et al., 2018). On the other hand, PISA data show many students have a relatively good knowledge of the traditional and still essential aspects of reading - e.g., comprehending and recalling a text and writing a summary, but they still lack the relevant knowledge and skills to navigate a digital world - e.g., navigating non-linear and conflicting information from multiple source texts (OECD, 2021a). The message is clear: to become proficient readers in a digital world, students need a solid foundation of reading but also the ability to think critically and adjust their behavior according to the task and motivate themselves to persevere in the face of difficulties. The results are also valuable for assessment designers to identify important reading performance indicators and align with the new reading assessment framework in the digital environment such as the recently approved 2026 National Assessment of Educational Progress (NAEP) Reading Framework (National Assessment Governing Board, 2022).

There is evidence that learning with mobiles is more effective than traditional teaching methods that only use pen-and-paper or desktop computers (Sung et al., 2016). Likewise, classroom interventions aimed at developing students' assessment of information reliability have proven to improve students' critical thinking when comprehending multiple documents (Pérez et al., 2018). However, adding specific subjects on digital skills or spending more time using digital devices in school without adjusting other parts of the curriculum could be problematic. The challenge is to respond to the changing needs while minimizing the expansion and overload of content or the replacement of other activities that are also important. One possibility for balancing curricular updating and excess is focusing on the process of reading – with process data being its gold mine - rather than on the product of reading. This study highlights some practices that teachers could use as cross-cutting themes and competencies in classroom settings. For example, teaching students to spend more time reading online texts without rush transitions

may help mitigate gender and socio-economic differences and enhance students' performance when navigating multiple source and dynamic texts.

In this paper we also showcased a novel method to cluster the navigation patterns by computing the distance similarity measure by the DTW algorithm. Specifically, we introduced two sequential indicators of navigation activity: the navigation sequence that tracks the page transition path and the time sequence that records the time duration on each visited page and computed pairwise sequence distance similarity measures in sequence page transition and time allocation sequence respectively. As reviewed in the sequence distance similarity measures, there are many approaches that we could use to calculate the sequence distances.

The DTW method that we highlighted in this study features in identifying the optimal warping path between the two sequences, which could be more beneficial in analyzing the time-series process data that cares more about the sequence shapes across time windows (Kurbalija et al., 2011). The DTW algorithm can be considered as a generalization of Euclidian distance where it is not necessary that the i-th point of one time series must be aligned to the i-th point of the other time series. This method allows elastic shifting of the time axis where in some points time "warps". The DTW algorithm computes the distance by finding an optimal path in a matrix of distances between points of two-time series. The Euclidian distance can be seen as a special case of DTW where only the elements on the main diagonal of the matrix are taken into account. It would be interesting to explore the possibility of constructing a time-weighted sequence similarity measure in the future study to combine the two sequence clustering analyses in a one-time run.

DTW is a distance measurement that is robust to time series phase perturbations; however, its quadratic complexity greatly impairs the efficiency of time series clustering (Cai et

al., 2021). Time series clustering is a well-studied field and the clustering algorithms can roughly be categorized into four classes: hierarchical, model-based, partition-based and density-based time series clustering. Partition-based time series clustering that was used in this study partitions the dataset into clusters by minimizing the overall distances of time series to their respective cluster centers. Typically, the center of a cluster, e.g., in K-means based time series clustering (e.g., Huang et al., 2016; Smith et al., 2018), is regarded as the point-wise average of the contained time series; however, such centers may poorly represent the common temporal pattern when phase perturbation appears. This could also lead to an unstable initial value for the k-centers to replicate the patterns in clustering (Cai et al; 2021). Other methods have been proposed in the literature as being more suitable to identify more accurate cluster centers. For example, K-medoids, recommended by Kaufman and Rousseeuw (2009) and also used in this study, regards the time series that has the least sum of distances to other time series as the center of a cluster, which helps locate the optimal initial center of a cluster in a more controllable way. It is recommended to use the K-medoids clustering method based on the DTW distance measurement or use a global averaging method such as KDBA (Petitjean et al., 2011) to generate the centers that adapt to DTW distance.

The disparities found in navigation behaviors by gender and socio-economic status also called for attention. A reading score gap was found between boys and girls in certain navigation patterns. This raised a new question for the tailored instruction for different groups with the aim to advance their potentiality to achieve a higher score. For instance, the navigation patterns that showed boys did not achieve the same score level as girls could be highlighted to the boys' instructors to help them pinpoint the reasons for the lack and readjust their navigation behavior in future reading task completion.

In addition, three limitations in this study may merit a discussion. First, we showcased how the navigation patterns could be identified through a data-driven approach with only one item. The research results could be further validated by using more items in the future to examine whether student's navigation patterns are consistent across items. Second, the example reading task showcased in this study only involved three pages, which might not be rich enough to show all the features of the DTW algorithm. It would be interesting to apply the DTW method to relatively longer sequences, for instance, in web search applications and examine the navigation patterns with reading comprehension in the future study. Third, we computed the DTW distance of page transition sequences and time allocation sequences separately in this study. Thus, the differences of time spent on each page (e.g., reading, re-reading, pausing or skimming on a certain page) could not be effectively affiliated with the page transition sequences in the DTW distance computation in a general mode. It would be interesting to construct a time-weighted sequence similarity measure with DTW in the future study to combine both the page transition and time interval between pages.

Although our study relied on data from 69 country samples, we estimated and presented results that did not consider country specificities in navigation behaviors. The aim of our work was in fact to develop and present an approach to identify navigation patterns and apply this approach to illustrate its potential to characterize how 15-year-old students approach reading tasks and inequalities in emerging reading skills. Detailed analyses of heterogeneities across countries and reasons that could explain such differences are both relevant and important but were beyond the scope of this work. Exploring differences in the navigation behaviors of students in different countries, differences in how navigation behavior relates to reading proficiency in different countries, and between-country differences in gender and socio-

economic disparities in navigation should be explored in further research. Such work could shed light on the role of language in shaping between country differences as well as the different way in which students in different countries are socialized and taught to approach digital reading texts. The length of the same text differs in languages: for example, character-based languages often produce shorter text length than alphabetic languages (Yamamoto et al., 2018) and typing speed in character-based language with phonetic scripts such as Korean could be faster than other languages in completing open-ended questions (Ercikan et al., 2021). There is also evidence that collectivistic cultures promote contextual understanding of situations and think more holistically, directing attention to all the elements they are presented with in a given situation and to relationships between items (Nisbett, et al., 2001). By contrast, in Western cultures, personal autonomy and formal logic are emphasized and individuals operating in these cultures are more likely to process information analytically, directing attention to particular items, objects and categories (Chua et al., 2005).

Our study focuses on exploring different patterns of navigation behaviors among those students who engaged in navigation behavior when they were presented with the reading task used in this study. As indicated, around a third of students internationally engaged in navigation, 46.2% of students skipped the task without engaging with the material, and 20.1% didn't execute any form of navigation. It is impossible to know with precision the reasons why some students failed to engage with the assessment task and, therefore, excluded from this study. However, it is not rare that some students may lack motivation or feel disengaged when taking low-stakes assessments with no consequences for them (Goldhammer et al., 2014; Goldhammer et al., 2017; Lundgren & Eklöf, 2020; Borgonovi et al., 2021). Furthermore, test engagement tends to decline as the test progresses, tends to be lower among boys, and differs across countries (Balart &

Oosterveen, 2019; Borgonovi & Biecek, 2016). We suggest future studies expand the findings of the present research to develop comprehensive models of test-taking behaviors across all competency levels.

The research potential of process data in reading tasks can be maximized if efforts could be made to design items that allow for clear identification of different solution and navigation strategies (He et al., 2021); ideally, these different strategies would be mapped to cognitive theories that researchers might be interested to test with the aid of process data, thus better supporting students' learning and assessments.

Acknowledgement

The authors thank Miyako Ikea, Giannina Rech and the OECD Programme for International Student Assessment (PISA) team for granting access to the data source and instruments; Eugenio Gonzales, Irwin Kirsch, Miyako Ikeda and Giannina Rech for helpful discussions on the study aim and design; Michael Wagner and Mathew Kandathil for data extraction and preprocessing; the Center for Psychometrics, Statistics, and Data Science and the Center for Global Assessments at the Educational Testing Service for their support. All views expressed in this paper are solely those of the authors and do not necessarily reflect those of the OECD or its member countries, the British Academy, or Education Testing Service. 4JS. This project is partially supported by National Science Foundation grants IIS-1633353, ETS Research Funding and the British Academy Global Professorship scheme.

References

Amin, T. B., & Mahmood, I. (2008). Speech recognition using dynamic time warping. In *2008 2nd International Conference on Advances in Space Technologies* (pp. 74-79). IEEE.

Arya, M. S., Deepa, R., & Gandhi, J. (2021). Dynamic Time Warping-Based Technique for Predictive Analysis in Stock Market. In *Proceedings of 6th International Conference on Recent Trends in Computing* (pp. 23-36). Springer, Singapore.

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications, 10*, 3798. https://doi.org/10.1038/s41467-019-11691-y

Barzilai, S., A. Zohar and S. Mor-Hagani (2018). Promoting Integration of Multiple Texts: a Review of Instructional Approaches and Practices. *Educational Psychology Review, 30*(3), 973-999, http://dx.doi.org/10.1007/s10648-018-9436-8.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles, 42*(1–2), 1–21. http://doi.org/10.1023/A:1007051109754

Borgonovi, F. (2016). Video gaming and gender differences in digital and printed reading performance among 15-year-olds students in 26 countries. *Journal of Adolescence, 48,* 45-61.

Borgonovi, F. (2022). Is the literacy achievement of teenage boys poorer than that of teenage girls, or do estimates of gender gaps depend on the test? A comparison of PISA and PIAAC. *Journal of Educational Psychology, 114*(2), 239–256. https://doi.org/10.1037/edu0000659

Borgonovi, F., & Biecek, P. (2016). An international comparison of students' ability to endure

fatigue and maintain motivation during a low-stakes test. *Learning and Individual Differences, 49,* 128-137.

Borgonovi, F., Choi, A., & Paccagnella, M. (2021). The evolution of gender gaps in numeracy and literacy between childhood and young adulthood. *Economics of Education Review, 82*, 102119. https://doi.org/10.1016/j.econedurev.2021.102119

Borgonovi, F., & Pokropek, A. (2021a). The evolution of socio-economic disparities in literacy skills from age 15 to age 27 in 20 countries. *British Educational Research Journal, 47*(6), 1560-1586.

Borgonovi, F., & Pokropek, A. (2021b). The evolution of the association between ICT use and reading achievement in 28 countries. *Computers & Education Open,* 100047, https://doi.org/10.1016/j.caeo.2021.100047

Buchmann, C., DiPrete, T. A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology, 34*(1), 319–337. https://doi.org/10.1146/annurev.soc.34.040507.134719

Britt, M. & Rouet, J-F. (2012). Learning with Multiple Documents, in Kirby, J. and M. Lawson (eds.), *Enhancing the Quality of Learning*, Cambridge University Press, Cambridge, http://dx.doi.org/10.1017/cbo9781139048224.017.

Britt, M. A., Rouet, J.-F., & Durik, A. (2018). Representations and processes in multiple source use, in *Educational Psychology Handbook Series. Handbook of Multiple Source Use*. eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (London: Routledge), 17–33.

Cai, B., Huang, G., Samadiani, N., Li, G., & Chi, C. H. (2021). Efficient time series clustering by minimizing dynamic time warping utilization. *IEEE Access, 9*, 46589-46599.

Chang, P. C., Fan, C. Y., Lin, J. L., & Lin, J. J. (2008). Integrating a piecewise linear

representation method with dynamic time warping system for stock trading decision making. In *2008 Fourth International Conference on Natural Computation* (Vol. 2, pp. 434-438). IEEE.

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

Cho, B.-Y., Afflerbarch, P., & Han, H. (2018). Strategic processing in accessing, comprehending, and using multiple sources online, in *Educational Psychology Handbook Series. Handbook of Multiple Source Use*. eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (London: Routledge), 133–150.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *PNAS, 102*(35), 12629-12633.

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading, 42*(2), 288-325. http://dx.doi.org/10.1111/1467-9817.12269.

Coleman, J. S., Campbell, E. Q., Hobson, C. F., McPartland, A. M., Mood, A. M., Weinfield, F. D., et al. (1966). *Equality of educational opportunity.* U. S. Government Printing Office, Washington, DC.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

DiPrete, T. A., & Buchmann, C. (2013). *The Rise of Women : The Growing Gender Gap in Education and What it Means for American Schools.* New York, NY: Russell Sage

Foundation.

Delgado, P., Vargas, C., Ackerman, R., Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review, 25,* 23-38, http://dx.doi.org/10.1016/j.edurev.2018.09.003.

Dong, G., & Pei, J. (2007). *Sequence data mining* (Vol. 33). Springer Science & Business Media.

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

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational Assessment, 25*(3), 179-197.

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(1), 1–37. https://doi.org/10.18637/jss.v040.i04

Gao, Y., Cui, Y., Bulut, O., Zhai, X., & Chen, F. (2022). Examining adults' web navigation patterns in multi-layered hypertext environments. *Computers in Human Behavior, 129,* 107142.

Goldhammer, F., Naumann, J., & Keβel, Y. (2013). Assessing individual differences in basic computer skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment, 29*(4), 263–275.

Goldhammer, F., Naumann, J. Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill. *Journal of Educational Psychology, 106*(3), 608-626.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in

PIAAC. *OECD Education Working Papers, No. 133*, OECD Publishing, Paris, https://doi.org/10.1787/5jlzfl6fhxs2-en.

Goldhammer, F., Martens, T., and Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC. *Large-scale Assessments in Education, 5*, 18.

Guo, A., & Siegelmann, H. T. (2004). Time-Warped Longest Common Subsequence Algorithm for Music Retrieval. In *ISMIR*. https://groups.cs.umass.edu/binds/wp-content/uploads/sites/21/2019/05/2004_AnYuan_ISMIR.pdf

Hahnel, C., Goldhammer, F., Kroehne, U., & Naumann, J. (2018). The role of reading skills in the evaluation of online information gathered from search engine environments. *Computers in Human Behavior, 78*, 223–234. https://doi.org/10.1016/j. chb.2017.10.004

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal, 29*(2), 147–160.

Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology, 10*, 2461. doi:10.3389/fpsyg.2019.02461

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining, 7* (1), 33–50. doi:10.5281/ZENODO.3554705

He, Q., Borgonovi, F., & Paccagnella, M. (2019). *Using Process Data to Understand Adults' Problem-Solving Behaviour in the Programme for the International Assessment of Adult Competencies (PIAAC): Identifying Generalised Patterns Across Multiple Tasks with Sequence Mining.* OECD Publishing; OECD Education Working Papers. http://doi.org/10.1787/650918f2-en

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. *Computers & Education, 166,* 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., Liao, D., & Jiao, H. (2019). Clustering behavioral patterns using process data in PIAAC problem-solving items. In B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement* (pp. 189-212), Methodology of Educational Measurement and Assessment (book series), Springer.

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research* (pp. 173–190). Springer. https://doi.org/10.1007/978-3-319-19977-1_13

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

Herder, E. (2005). Characterizations of user web revisit behavior. In *Proceedings of Workshop on Adaptivity and User Modelling in Interactive Systems*, 32–37.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the Association for Computing Machinery, 18*, 341–343.

Hirschberg, D. S. (1977). Algorithms for the longest common subsequence problem. *Journal of the Association for Computing Machinery, 24*(4), 664–675.

Huang, X., Ye, Y., Xiong, L., Lau, R. Y., Jiang, N., & Wang, S. (2016). Time series k-means: A

new k-means type smooth subspace clustering for time series data. *Information Sciences, 367*, 1-13.

Jenkins, C., Corritore, C. L., & Wiedenbeck, S. (2003). Patterns of information seeking on the web: A qualitative study of domain expertise and web expertise. *IT & Society, 1* (3), 64–89.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Pearson Prentice Hall.

Kaufman, L., & Rousseeuw, P. J. (1990). *Partitioning Around Medoids (Program PAM)*, Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 68–125, doi:10.1002/9780470316801.ch2

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining, 5*(1), 190–219.

Kurbalija, V., Radovanović, M., Geler, Z., & Ivanović, M. (2011). The influence of global constraints on DTW and LCS similarity measures for time-series databases. In *Third International Conference on Software, Services and Semantic Technologies S3T 2011* (pp. 67-74). Springer, Berlin, Heidelberg.

Lawless, K. A., & Kulikowich, J. M. (1996). Understanding Hypertext Navigation through Cluster Analysis. *Journal of Educational Computing Research*, *14*(4), 385–399. http://doi.org/10.2190/DVAP-DE23-3XMV-9MXH

Lee, H. (2007). Instructional design of web-based simulations for learners with different levels of spatial ability. *Instructional Science, 35*, 467-479.

Levenshtein, V. I. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission, 1*(1), 8–17.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady, 10*, 707–710.

Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: An investigation of U.S. adults' employment status in PIAAC. *Frontiers in Psychology, 10*: 646. doi:10.3389/fpsyg.2019.00646

List, A. & P. Alexander (2018). Toward an Integrated Framework of Multiple Text Use. *Educational Psychologist, 54*(1), 20-39. http://dx.doi.org/10.1080/00461520.2018.1505514.

Liu, J., Zhang, S., & Yang, J. (2004). Characterizing web usage regularities with information foraging agents. *IEEE Transactions on Knowledge and Data Engineering, 16* (5), 566–584. https://doi.org/10.1109/TKDE.2004.1277818

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation, 26*(5-6), 275-301, https://doi.org/10.1080/13803611.2021.1963940

Magliano, J. P., McCrudden, M. T., Rouet, J. F., & Sabatini, J. (2017). The modern reader: Should changes to how we read affect research and theory? In *The Routledge handbook of discourse processes* (pp. 343-361). Routledge.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(1),11–48.

National Assessment Governing Board. (2022). *Reading Framework for the 2026 national*

Assessment of Educational Progress. U.S. Department of Education. https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/reading/2026-reading-framework/naep-2026-reading-framework.pdf

Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior, 53,* 263–277. https://doi.org/10.1016/j.chb.2015.06.051

Naumann J. (2019). The Skilled, the Knowledgeable, and the Motivated: Investigating the Strategic Allocation of Time on Task in a Computer-Based Assessment. *Frontiers in psychology*, *10,* 1429. https://doi.org/10.3389/fpsyg.2019.01429

Naumann, J., & Salmerón, L. (2016). Does Navigation Always Predict Performance? Effects of Navigation on Digital Reading are Moderated by Comprehension Skills. *The International Review of Research in Open and Distributed Learning, 17*(1). https://doi.org/10.19173/irrodl.v17i1.2113

Nisbett, R.E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review, 108*(2): 291-310. https://doi.org/10.1037/0033-295x.108.2.291

Notten, N., Peter, J., Kraaykamp, G., Valkenburg P.M. (2009). Digital divide across borders—A cross-national study of adolescents' use of digital technologies. *European Sociological Review, 25*(5), 551-560.

OECD (2018). *Effective Teacher Policies: Insights from PISA.* Paris: OECD Publishing. https://doi.org/10.1787/9789264301603-en.

OECD (2011). *PISA 2009 Results: Students on Line. Digital Technologies and Performance* (Volume VI), OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264112995-en.

OECD (2015a). *In it together: Why less inequality benefits all.* Paris: OECD Publishing.

OECD. (2015b). *The ABC of Gender Equality in Education.* Paris: OECD Publishing.

OECD (2019a). *PISA 2018 Released Field Trial and Main Survey New Reading Items*. OECD Publishing, Paris, https://www.oecd.org/pisa/test/PISA2018_Released_REA_Items_12112019.pdf

OECD (2019b). *PISA 2018 Results (Volume II): Where All Students Can Succeed.* OECD Publishing, Paris, https://doi.org/10.1787/b5fd1b8f-en

OECD (2021a). *21st-Century Readers: Developing Literacy Skills in a Digital World, PISA,* OECD Publishing, Paris, https://doi.org/10.1787/a83d84cb-en.

OECD (2021b). *OECD Skills Outlook 2021: Learning for Life*. OECD Publishing, Paris, https://doi.org/10.1787/0ae365b4-en.

Permanasari, Y., Harahap, E. H., & Ali, E. P. (2019). Speech recognition using dynamic time warping (DTW). In *Journal of Physics: Conference Series* (Vol. 1366, No. 1, p. 012091). IOP Publishing.

Pérez, A., Potockia, A., Stadtler, M., Macedo-Rouet, M., Paul, J., Salmerón, L. Rouet, J.F. (2018). Fostering teenagers' assessment of information reliability: Effects of a classroom intervention focused on critical source dimensions. *Learning and Instruction, 58,* 53-64. http://dx.doi.org/10.1016/j.learninstruc.2018.04.006

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, *44*(3), 678-693.

Pokropek, A., Borgonovi, F. & Jakubowski, M. (2015). Socio-economic disparities in academic achievement: A comparative analysis of mechanisms and pathways. *Learning and Individual Differences, 42,* 10-18.

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

Ren, Z., Fan, C., & Ming, Y. (2016, November). Music retrieval based on rhythm content and dynamic time warping method. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)* (pp. 989-992). IEEE.

Richter, T., and Maier, J. (2017). Comprehension of multiple documents with conflicting information: a two-step model of validation. *Educational Psychologist, 52,* 148–166. doi: 10.1080/00461520.2017.1322968

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics, 20,* 53–65. doi:10.1016/0377-0427(87)90125-7.

Rouet, J. F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. *Text relevance and learning from text,* 19-52.

Saez, E., & Zucman, G. (2016). Wealth inequality in the United States since 1913: Evidence from Capitalized income tax data. *The Quarterly Journal of Economics, 131*(2), 519-578. http://dx.doi.org/10.1093/qje/qjw004.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing, 26*(1), 43-49.

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,* 1–20. doi:10.1186/s40536-020-00085-y

Salmerón, L., & García, V. (2011). Reading skills and children's navigation strategies in hypertext. *Computers in Human Behavior, 27*(3), 1143–1151.

http://doi.org/10.1016/j.chb.2010.12.008

Salmerón, L., Strømsø, H. I., Kammerer, Y., Stadtler, M., & van den Broek, P. (2018). Comprehension processes in digital reading. In M. Barzillai, J. Thomson, S. Schroeder, & P. van den Broek (Eds.), *Learning to read in a digital world* (pp. 91–120). John Benjamins.

Smith, K. E., Williams, P., Bryan, K. J., Solomon, M., Ble, M., & Haber, R. (2018, July). Shepard interpolation neural networks with k-means: a shallow learning method for time series classification. In 2018 *International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE.

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research 1990–2000. *Review of Educational Research, 75*(3), 417-453.

Solheim, O. J., & Lundetræ, K. (2018). Can test construction account for varying gender differences in international reading achievement tests of children, adolescents and young adults? – A study based on Nordic results in PIRLS, PISA and PIAAC. *Assessment in Education: Principles, Policy & Practice, 25*(1), 107-126.

Studer, M., & Ritschard, G. (2014). A comparative review of sequence dissimilarity measures. *LIVES Working Papers, 33*, 1–47. https://doi.org/10.12682/lives.2296-1658.2014.33

Suárez-Álvarez, J., R. Fernández-Alonso and J. Muñiz (2014). Self-concept, motivation, expectations, and socioeconomic level as predictors of academic performance in mathematics. *Learning and Individual Differences, 30*, 118-123. http://dx.doi.org/10.1016/j.lindif.2013.10.019.

Sukkarieh, J. Z., von Davier, M., & Yamamoto, K. (2012). From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks. *Research Report*

No. RR-12-25. Educational Testing Service.

Sung, Y., K. Chang and T. Liu (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252-275. http://dx.doi.org/10.1016/j.compedu.2015.11.008.

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

Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, *47*(1), 3-35. doi:10.3102/10769986211010467

Ulitzsch, E., He, Q., Ulitzsch, V., Nichterlein, A., Molter, H., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika*, 1–25. doi:10.1007/s11336-020-09743-0

Van Meter, P., List, A., Lombardi, D., Kendeou, P. (eds.) (2020). *Handbook of Learning from Multiple Representations and Perspectives*. Routledge. http://dx.doi.org/10.4324/9780429443961.

Van de Werfhorst, H. G., & Mijs, J.B. (2010). Achievement Inequality and the Institutional Structure of Educational Systems: A Comparative Perspective. *Annual Review of Sociology, 36*, 407-428.

Vaughan, M., & Dillon, A. (2006). Why structure and genre matter to users of digital information: a longitudinal study with readers of a web-based newspaper. *International Journal of Human-Computer Studies, 64*, 502–526.

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

Voyer, D., & Voyer, S. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174-1204.

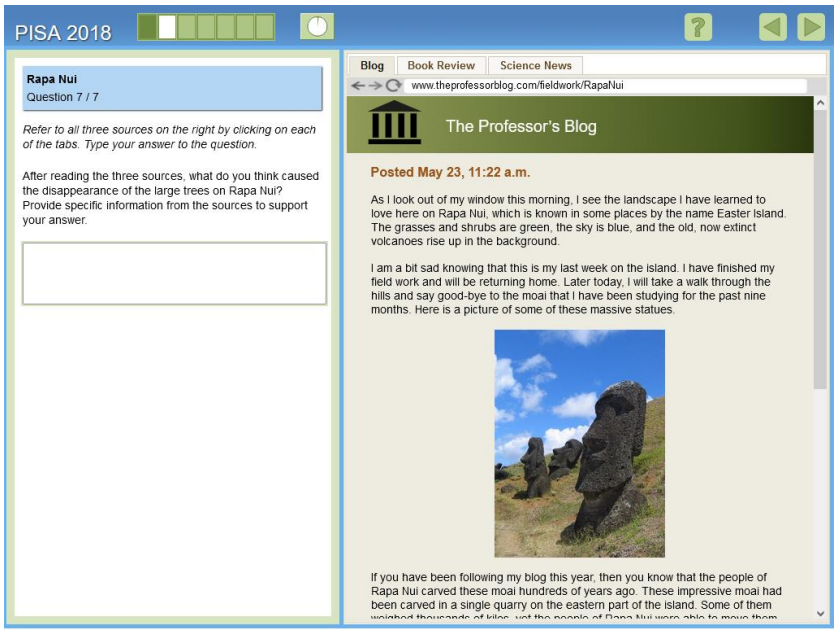
Willingham, W. W., & Cole, N. S. (2013). *Gender and fair assessment.* Routledge.

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2018). Development and implementation of a machine-supported coding system for constructed-response items in PISA. *Psychological Test and Assessment Modeling, 60*(2), 145–164.

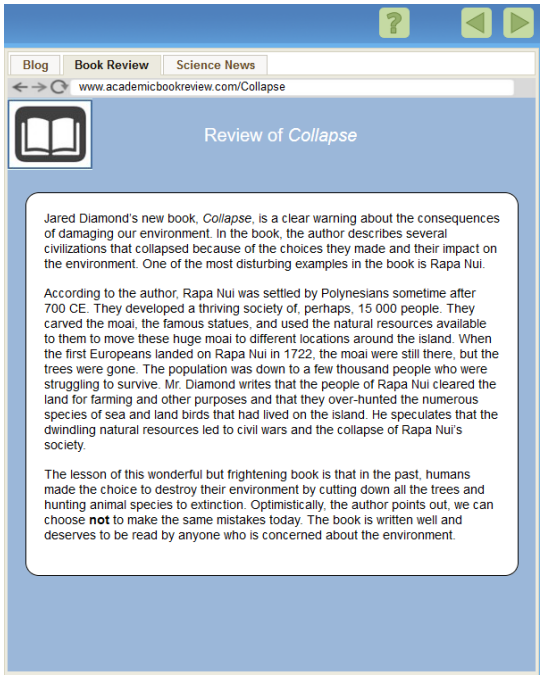
Zhou, M. (2014). Gender differences in web search perceptions and behavior: Does it vary by task performance? *Computers & Education, 78*, 174-184.

Figure 1

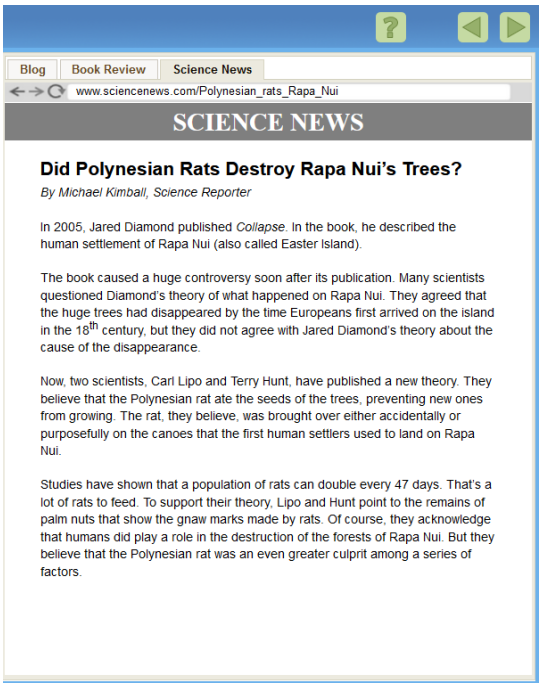
Screenshots of CR551Q11 homepage (Blog page), Book Review page and Science News pages



Homepage (Blog Page)



Book Review Page



Science News Page

Note. Only the right part of screenshots on the Book Review page and the Science News page are shown, given the space limitation. The left side of the screen keeps the same as the homepage.
Resource: https://pisa2018-questions.oecd.org/platform/index.html?user=&domain=REA&unit=R551-RapaNui&lang=eng-ZZZ

Figure 2

An example of distance matrix computed by dynamic timing warping method

A	**5**	10	6	3	1	2	4	3
	4	6	3	1	0	1	3	3
	3	3	1	0	1	1	2	4
	2	1	0	1	3	4	4	7
	1	0	1	3	6	8	9	13
		1	**2**	**3**	**4**	**3**	**2**	**5**
		B						

Note. This example is to compute the distance similarity measure between two sequences: $A = \{1,2,3,4,5\}$ and $B = \{1,2,3,4,3,2,5\}$. The highlighted yellow path is the shortest path starting from the upper right diagonal corner to the lower left diagonal corner. The sum of value along the highlighted cells is the similarity score between sequences A and B.

Figure 3
Silhouette index in DTW sequence clustering for page transition and time on page sequences

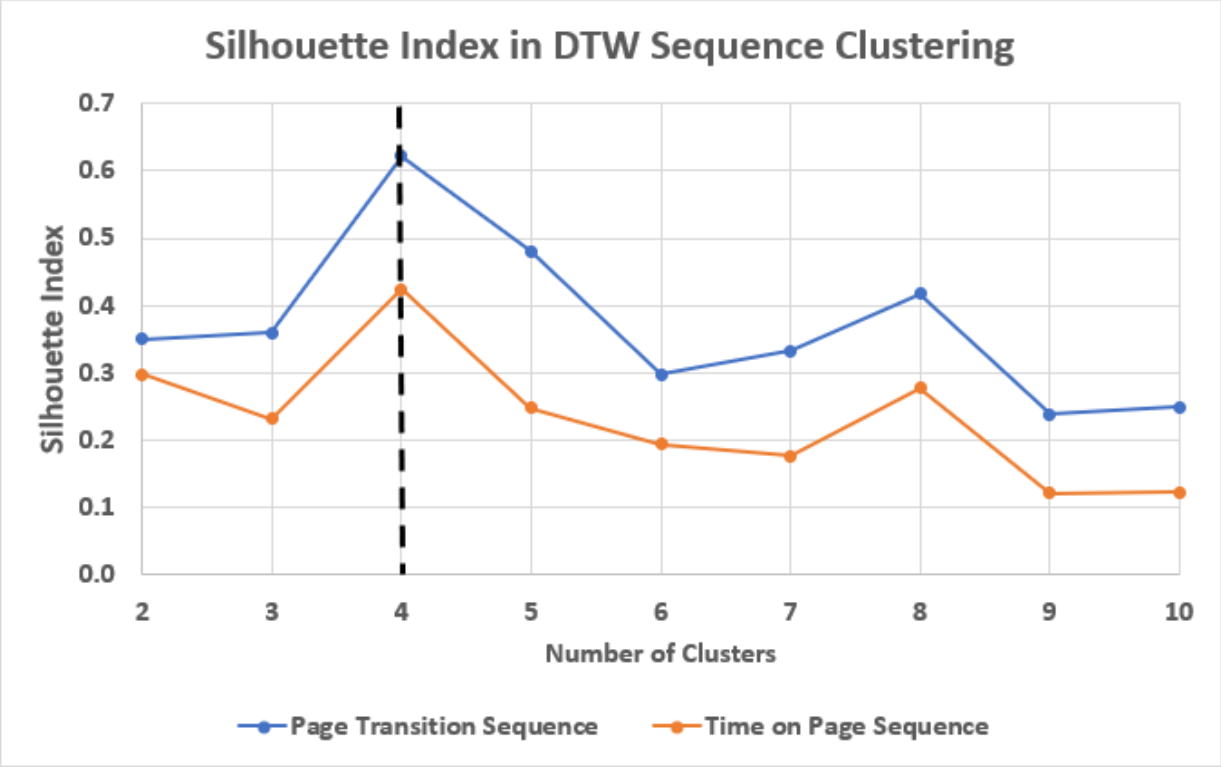
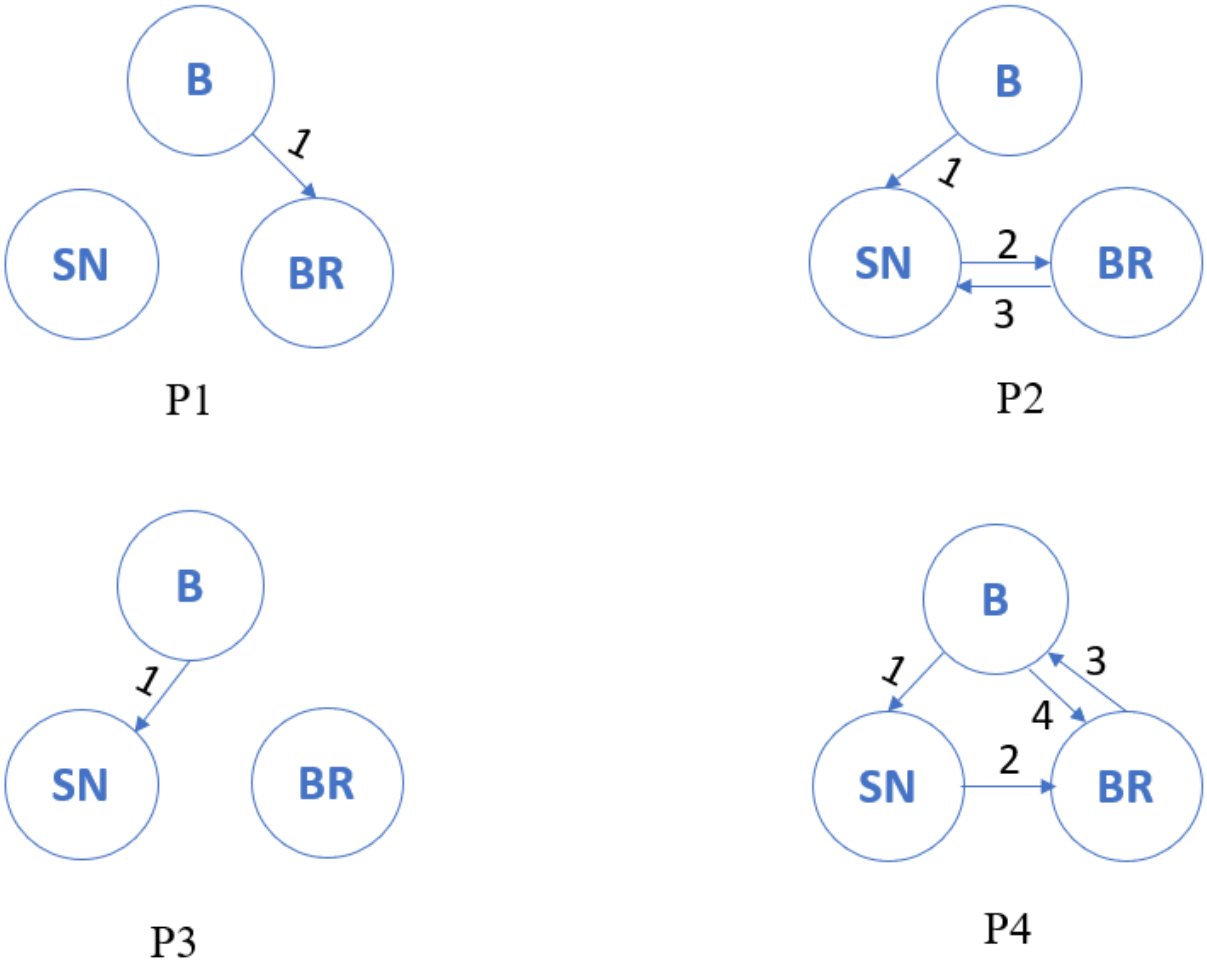


Figure 4

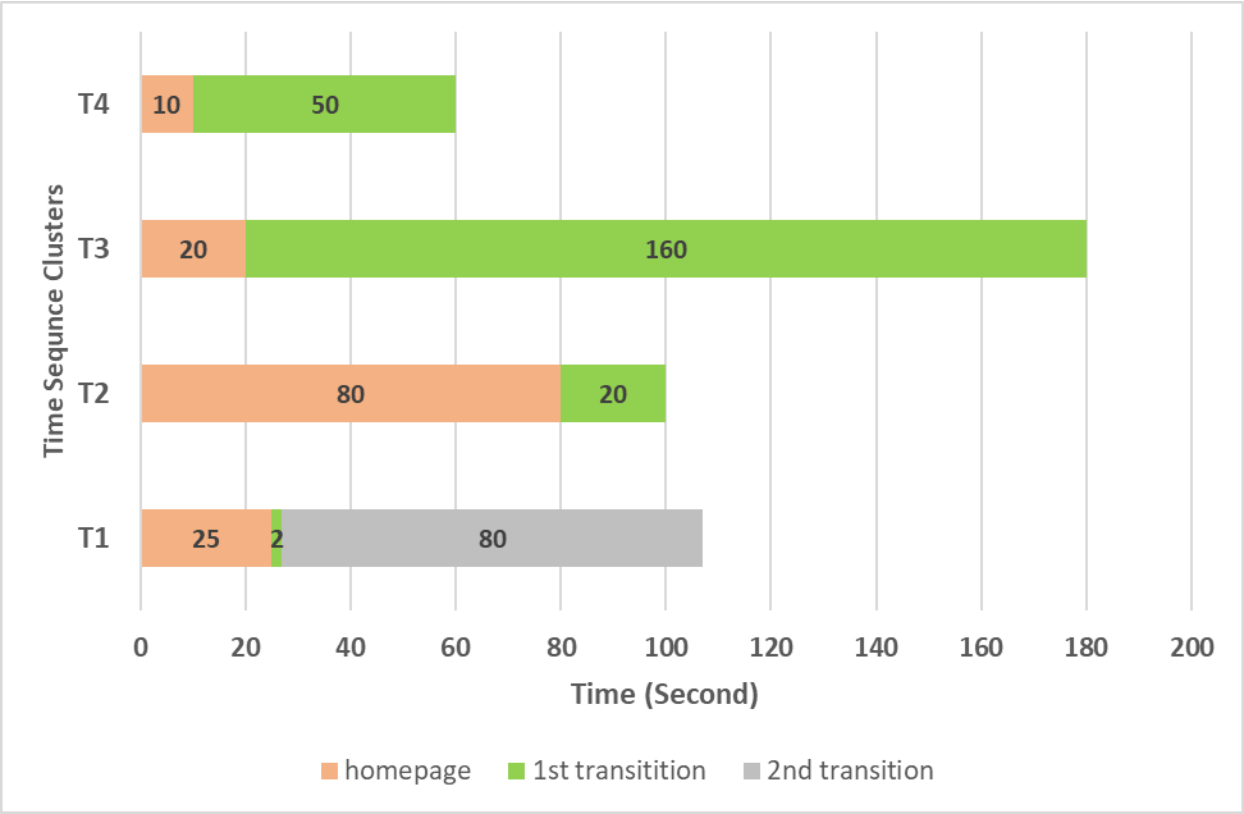
Four typical navigation page transit patterns derived from clustering analysis based on dynamic time warping distance similarity measure



Note. Circles represent the transition pages. Arrows represent the direction of transition. The numbers on each arrow represent the order of transition. B=Blog, BR=Book Review, SN=Science News.

Figure 5

Four typical sequences of time spent on transition pages derived from clustering analysis based on dynamic time warping distance similarity measure



Note. Each colored block represents the average amount of time spent on that page.

Figure 6

Association between reading performance and navigation patterns with joint membership in page and transition time clusters

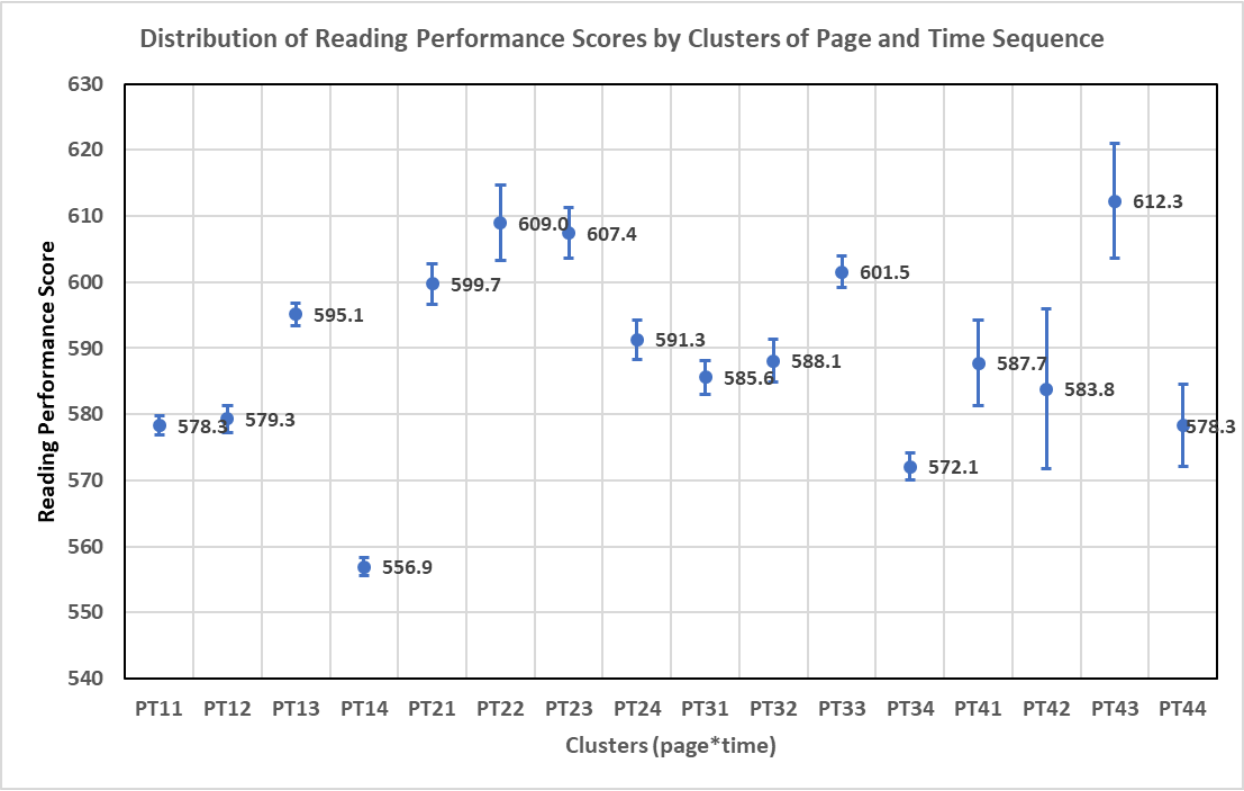


Figure 7

Gender disparities in reading score by navigation patterns

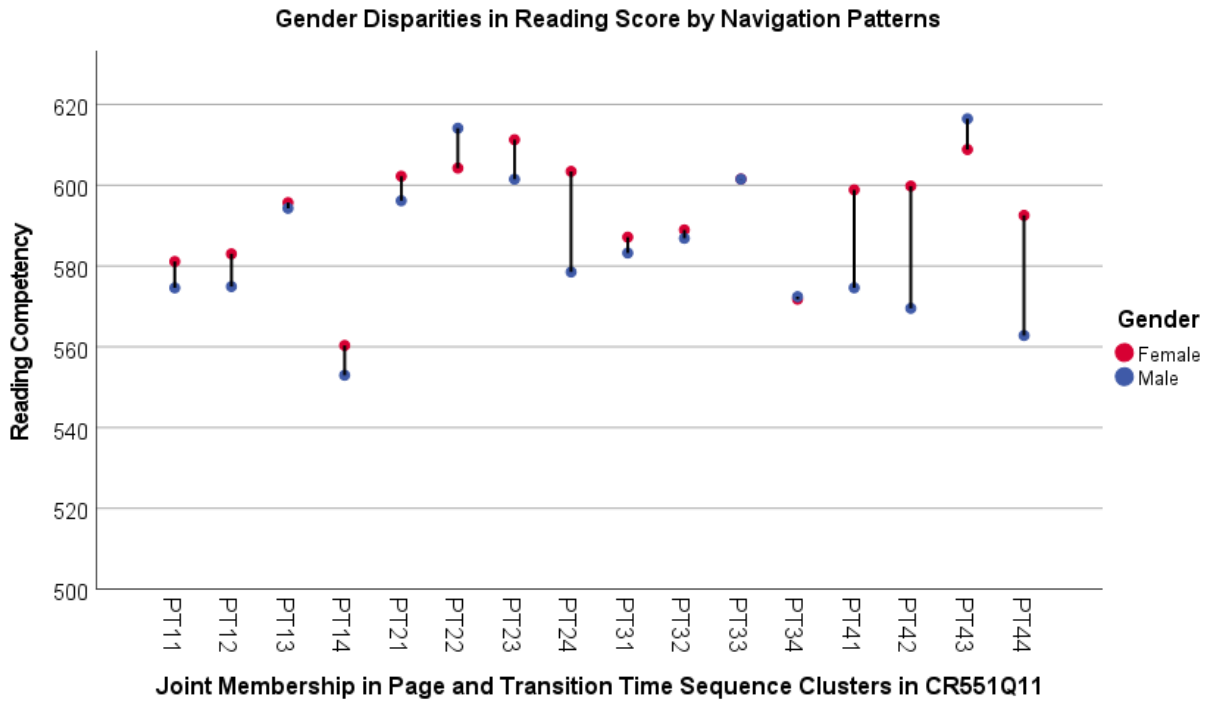


Table 1. Descriptive statistics of reading proficiency score, gender and socio-economic index by navigation patterns

Clusters	Sample Size	Percentage	Reading Proficiency Score			Gender		Socio-Economic Index		
			Mean	S.D.	S.E.	Girl (%)	Boy (%)	Mean	S.D.	S.E.
P1	11091	65.4%	574.13	82.05	0.78	65.2%	66%	0.22	0.94	0.01
P2	1952	11.5%	599.06	76.97	1.74	11.4%	12%	0.29	0.90	0.02
P3	3438	20.8%	585.07	72.38	1.23	20.8%	20%	0.28	0.92	0.02
P4	476	2.6%	587.75	85.03	3.90	2.6%	3%	0.28	0.92	0.04
T1	4651	27.4%	582.72	78.12	1.15	28.1%	27%	0.27	0.93	0.01
T2	2385	13.6%	583.75	79.43	1.63	13.6%	15%	0.33	0.92	0.02
T3	3675	23.5%	598.36	77.17	1.27	23.5%	19%	0.31	0.91	0.02
T4	6246	34.8%	564.66	80.83	1.02	34.8%	39%	0.16	0.94	0.01
PT11	3065	18.1%	578.31	79.53	1.44	18.3%	18%	0.24	0.95	0.02
PT12	1610	9.5%	579.26	80.29	2.00	9.1%	10.0%	0.31	0.93	0.02
PT13	2335	13.8%	595.14	79.78	1.65	14.7%	12.6%	0.30	0.91	0.02
PT14	4081	24.1%	556.94	82.43	1.29	23.1%	25.3%	0.14	0.95	0.01
PT21	610	3.6%	599.73	74.19	3.00	3.8%	3.4%	0.35	0.87	0.04
PT22	196	1.2%	608.99	78.95	5.64	1.1%	1.3%	0.37	0.86	0.06
PT23	405	2.4%	607.44	75.69	3.76	2.6%	2.1%	0.33	0.88	0.04
PT24	741	4.4%	591.30	78.72	2.89	4.0%	4.9%	0.19	0.93	0.03
PT31	819	4.8%	585.57	72.62	2.54	5.1%	4.4%	0.31	0.88	0.03
PT32	526	3.1%	588.07	73.78	3.22	3.2%	3.0%	0.35	0.92	0.04
PT33	855	5.0%	601.53	69.67	2.38	5.7%	4.1%	0.31	0.93	0.03
PT34	1238	7.3%	572.10	71.06	2.02	6.7%	8.1%	0.22	0.93	0.03
PT41	157	0.9%	587.75	83.43	6.66	0.9%	1.0%	0.32	0.97	0.08
PT42	53	0.3%	583.81	90.80	12.47	0.3%	0.4%	0.39	0.75	0.10
PT43	80	0.5%	612.28	78.20	8.74	0.5%	0.5%	0.50	0.82	0.09
PT44	186	1.1%	578.32	86.09	6.31	1.0%	1.2%	0.12	0.93	0.07
Total	16957	100.0%	579.60	80.13	0.62	100%	100%	0.25	0.93	0.01

Note. S.E. refers to standard error. S.D. refers to standard deviation.