Within-item response processes as indicators of test-taking effort and motivation

Erik Lundgren and Hanna Eklöf

Department of Applied Educational Science, Umeå University, Umeå, Sweden

ABSTRACT

The present study used process data from a computer-based problem-solving task as indications of behavioural level of test-taking effort, and explored how behavioural item-level effort related to overall test performance and self-reported effort. Variables were extracted from raw process data and clustered. Four distinct clusters were obtained and characterised as high effort, medium effort, low effort, and planner. Regression modelling indicated that among students that failed to solve the task, level of effort invested before giving up positively predicted overall test performance. Among students that solved the task, level of effort was instead weakly negatively related to test performance. A low level of behavioural effort before giving up the task was also related to lower self-reported effort. Results suggest that effort invested before giving up provides information about test-takers’ motivation to spend effort on the test. We conclude that process data could augment existing methods of assessing test-taking effort.

KEYWORDS

Process data; test-taking effort; test-taking motivation; problem solving; PISA 2012; computer-based assessment

Introduction

Items in large-scale assessments are often scored by assigning an integer value of 0 or 1 to a fixed or open response. Less is known about how a participant arrived at the score. With computer-based assessments comes the possibility to trace human–computer interactions, which can give a glimpse of what was done between item presentation and answer selection. These digital traces have been referred to as a kind of process data, and can be used as a source of validity evidence based on response processes (American Educational Research Association et al., 2014; Goldhammer et al., 2017). In previous studies, response process data have been used to understand students’ plans, skills, and misconceptions (LaMar, 2014; Rafferty et al., 2020), in real-time tracing of skills (Polyak et al., 2017), to identify problem-solving strategies
(Greiff et al., 2015; He, Borgonovi, & Paccagnella, 2019; Liu et al., 2018), and to assess science enquiry skills (Buckley et al., 2010; Gobert et al., 2013). One area that has received less attention from research on process data is test-taking effort, which we aim to explore in this study by examining test-takers’ within-item behaviours in a Programme for International Student Assessment (PISA) 2012 problem-solving item.

Test-taking effort

International large-scale assessments such as PISA or Trends in International Mathematics and Science Study (TIMMS) are considered low-stakes assessments for the participating students – no grades are affected, and neither score nor feedback returned to the students. Thus, test-takers might not be incentivised to perform their best, which could threaten the validity of the assessment. A review by Wise and DeMars (2005) indicates that level of test-taking motivation and test performance are positively related, and reported group effect sizes ranging from −0.04 to 1.49, and correlation effect sizes between 0.23 to 0.38. The positive relationship between motivation and performance has been corroborated by later studies (Cole et al., 2008; Eklöf & Knekta, 2017).

Since there are different understandings of the concepts of test-taking effort, motivation, and engagement, and since these terms at times have been considered synonymous (Wise & Gao, 2017), we will attempt to be clear on how we understand the terms. We understand general motivation to be a two-component latent construct, where one component refers to a direction towards some desired goal state, and the other component indicates the magnitude of the subjective value of attaining the goal. The value component is similar to attainment value in the expectancy-value model of achievement motivation (Wigfield & Eccles, 2000), and similar to what Atkinson (1964) means by strength motivation or strength of tendency. We assume that the value component will determine the maximum amount of resources an agent would be willing to spend to attain the goal state. In line with the previous general definition of motivation, test-taking motivation would then be a specific motivation that has a fixed direction towards trying to maximise the score on a test, and that test-taker’s level of motivation will regulate the maximum amount of effort they are willing to spend on pursuing the test-taking goal. The term test-taking effort can then be understood as the amount of resources that a test-taker uses to try to achieve the best possible score on a specific test. Test-taking effort is thus understood essentially in the same way as how Wise and Demars (2005) define it: “test-taking effort as a student’s engagement and expenditure of energy toward the goal of attaining the highest possible score on the test” (p. 2). We assume engagement in this context to simply mean that a test-taker actively tries to solve the test items in the test, and whenever this is not the case, then test-takers are disengaged. We would rather use the
term “resources” instead of “energy”. Even if actual physical energy use is perhaps the ideal measurement of effort, it seems difficult to measure reliably. Instead, we understand resources as anything that a student is using to try to solve a test item, for example, behaviours and cognitions, which happen over time. Further, some students will probably be more efficient in solving tasks, where efficiency means that less effort and fewer resources had to be used to achieve the goal of a task. For example, if two test-takers solve a test task, one might easily solve the task (need to use fewer resources) and the other may need to use great amounts of resources. In this comparison, the difference in amount of effort would primarily indicate a difference in efficiency, and less could be inferred about differences in their underlying level of test-taking motivation, only that both students were at least willing to invest the amount of resources they needed to solve the item. Now, consider a case comparing two test-takers that used different amounts of effort before giving up trying to solve a task while knowing that they had not found the right answer. The effort that these test-takers spent reveals the maximum amount of resources that they were willing to invest in trying to solve the item. Here, amount of effort is regulated by their level of motivation, so in this case we can infer that differences in effort reflect differences in underlying levels of motivation. What this argument boils down to is that observed test-taking effort may not always be equal to test-taking motivation. Rather, we would argue that a more nuanced definition and interpretation may be needed, something that process data could possibly help us achieve.

*Methods of assessing test-taking effort*

Test-taking effort has often been estimated by post-assessment self-reports. Self-reports have their strengths in terms of tapping subjective information, but might be affected by social desirability, and other psychological factors. If self-report items are administered close to the end of an assessment, they could also be affected by imperfect memory retrievals, leading to biased inferences from participants (Schwarz, 2007). Studies investigating the relationship between self-reported test-taking effort and test performance have reported positive correlations around 0.5 (Cole et al., 2008), between 0.16 and 0.35 (Eklöf & Knekta, 2017), and 0.25 (Wise & Gao, 2017). A meta-analysis by Silm et al. (2020) estimated the average correlation to be 0.33 (0.30–0.37).

In addition to self-reports, response times have been used as indicators of test-taking effort. Wise and Kong (2005) have proposed a measure of response-time effort (RTE), which is defined as the proportion of response times classified as solution behaviours, whereas complementary responses are considered to be rapid-guessing behaviour indicative of low test-taking effort. The classification depends on item-specific thresholds set by the researcher. Thresholds have been specified by setting an omnibus 3-second threshold, by visual inspection of distributions for multimodal spikes, mixture-models (Kong et al., 2007), and
by using 10% of the average item response time, or 10 seconds if the average time was over 100 seconds (Wise & Gao, 2017). A meta-analysis (Silm et al., 2020) of the effect of RTE on test performance estimated a correlation of 0.72 (0.67–0.77). As researchers have taken an interest in response time measures of effort, there has at the same time been a call for a more detailed understanding of how good response times really are as proxies for careless responses (Rios et al., 2017) and disengaged responses that might not occur rapidly (Wise et al., 2020).

An example of previous research identifying effort-related constructs from within-task response process data is Gobert et al. (2015), who developed a detector focused on behaviours that indicated students that were disengaged from task goal; see Cocea and Weibelzahl (2009) for a similar example in the context of a web-based tutor, and Karumbaiah et al. (2018) and Mills et al. (2014), for detection of quitting behaviours in a learning game.

One clear benefit of assessing from response process data is that it is an unobtrusive, or stealth, assessment (Shute et al., 2016) that does not disrupt students’ work processes. Evidence from response processes could also be advantageous to self-report in cases where participants are not interested in answering questionnaires. Compared to the use of response time alone, the fine-grained within-item process data describe which actions were performed at a specific point in time. For a digital trace to be of interest, it must be informationally rich. Item designs that engage students in human–computer interaction where many different actions are possible and many different states can be reached, should increase the variability of possible behaviours. An item which is suitable to explore, with respect to identifying effort from response processes, is a problem-solving item from PISA 2012 called “Traffic” (Organisation for Economic Co-operation and Development [OECD], 2014a). In this item, students searched for the shortest route in a road network by repeated clicking. Test-takers could determine themselves whether the correct answer to the item was found, and do multiple trial and error attempts, making it possible to understand how much work was put into trying to solve the task.

**Research aims**

The overall aim of this study was to investigate if and how information from test-takers’ within-item response processes can advance our understanding of test-taking effort and test-taking motivation. We did this by extracting patterns of behaviour indicating different levels of effort, and by exploring how the behavioural effort indicators related to self-reported effort and performance on the PISA 2012 test. If participants’ within-item behavioural effort is indicative of general test-taking effort, which, in turn, affects performance, then, differences in within-item effort should be related to differences in in global performance in PISA. Further, if self-reported effort and behavioural effort are different
aspects of the same construct, they should be empirically related (see, however, Silm et al., 2019). The aims of the research could be described by the following statements:

1. Cluster test-takers’ within-item behaviours and interpret the clusters primarily with respect to level of effort.
2. Investigate if and how test performance varies by level of behavioural effort.
3. Investigate if and how self-reported test-taking effort varies with respect to behavioural effort.

**Method**

**Participants**

The sample consisted of 3,231 fifteen-year-old students from Denmark, Finland, Norway, and Sweden, who all took the computer-based part of the PISA 2012 assessment, in which the analysed Traffic item was part (see Table 1). The selected countries are largely a sample of convenience and are chosen on the basis of geographical, socioeconomic, educational, and cultural similarities in order to obtain a reasonable sample size while not including overly disparate subsamples.

**Instrumentation**

*Traffic problem-solving item*

In order to explore our research questions, we used log-file data from one problem-solving item. The item was a route-finding problem set in a scenario called Traffic (see Figure 1). The item was presented in the PISA 2012 assessment in a 40-minute block containing domains of computer-based mathematics, reading, and problem solving.

The item stimulus presents a schematic map of a road network together with an instruction to find the route with the shortest travel time between two fictional cities named Diamond and Einstein. The instructions included information that the shortest route would take 31 minutes. With this information,

| Country   | Test performance | Effort rating | Traffic task performance |
|-----------|------------------|---------------|-------------------------|
|           | n | M (SD) | n |Md (IQR) | n | % correct answer |
| Denmark   | 654 | 490 (88) | 601 | 8 (7–9) | 654 | 74 |
| Finland   | 1,729 | 508 (86) | 1656 | 8 (7–9) | 1,729 | 77 |
| Norway    | 415 | 482 (90) | 382 | 8 (6–9) | 415 | 75 |
| Sweden    | 433 | 486 (88) | 409 | 8 (6–8) | 433 | 71 |

*Note: Test performance shows mean (M) PISA math plausible values with standard deviation (SD) within parentheses. Effort rating shows median (Md) self-reported effort rated on a scale from 1–10 with inter-quartile range (IQR) within parentheses. Traffic task performance shows percentages of correct scores on the traffic item. The n shows the sample sizes.*
test-takers could determine if the route they had selected was the correct response to the item. The road network can be described as a graph with 16 nodes (cities) connected by 23 edges (roads), where each edge is labelled with a number indicating the travel time between cities (see Figure 1). Students could select a route by clicking on edges, which then highlighted them and

**Figure 1.** Relationships between the trace log file (1), events in the Traffic item (2), and extracted variables (3).

Note: A test-taker starts the item on Row 1, indicated by the value START ITEM in the event column. Rows 2 to 5 indicate four clicks on the time-window as annotated by arrow (a). From Rows 6 to 17, the participant clicked on path segments that added up to a complete route between Diamond and Einstein, as annotated by arrows (b). The test-taker ended the item as indicated at Row 18 and as annotated by arrow (c). Arrow (d) leads to a printout of the extracted variables (3) which summarises this test-takers’ behaviours. TOT = time on task; T2FA = time to first action; T2FR = time to first route; NOA = number of actions; RWR = repeated wrong routes; UNIQ_R = unique routes; APM = actions per minute.
accumulated the total travel time, displayed in a “total time” window in the lower left region of the display. This feature could be used as a quick check to see if a selected route fulfilled the time constraint of 31 minutes. To try another route, the student had to remove the selected route, either by clicking a reset button that cleared all the selected edges at once, or by de-selecting already highlighted edges by clicking them one by one.

The Traffic item is relatively easy to solve (see Table 1), and is knowledge lean, as the task for the students is to simply find a route that takes 31 minutes. It further provides opportunities to make repeated attempts and has a design that invites test-takers to do many clicks. These properties make the item suitable for observing test-taking behaviours in terms of effort and motivation. The item is available and can be tested at http://www.oecd.org/pisa/test-2012/testquestions/question2/.

Behavioral indicators of effort

Description of process data log file. The log file accompanying the PISA 2012 traffic item (see Figure 1) was a table with seven columns: cnt; schoolid; StIDStd; event; time; event_number, and event_value. The first three columns refer to Country, school ID, and student ID. The forth column, event, contains higher level descriptions of a within-item event, such as START_ITEM, click, dbclick (double click), ACER_EVENT, END_ITEM, which are associated with a timestamp in the fifth column, an event number in the sixth column, and a more detailed event value in the seventh column, such as hit_Diamondnowhere, which indicates that this road was clicked. The event_value associated with ACER_EVENT is a Boolean vector that describes the state of highlighted roads. The clicks and their associated event value tell which graphical element of the item was clicked at a given point in time. Figure 1 illustrates the relationship between the log file, item stimulus, and extracted variables.

Extracted variables. The logged events presented in Figure 1 cannot be used in clustering in their raw format. Therefore, the data were aggregated into seven variables describing different aspects of test-taking behaviour considered relevant to assess levels of effort in the traffic item.

Time on task (TOT) indicates the total time in seconds from start to end of the item. In combination with other variables, time on task could be indicative of effective solutions, the amount of effort, or reveal idle participants.

Time to first action (T2FA) indicates the time in seconds elapsed from item start to the first logged action. This could be an indicator of the time it took to read instructions and formulate a hypothesis about which path to explore. Since this item might be sensitive to quick non-intentional or exploratory clicks, a similar variable of time to first route (T2FR) was created to indicate how much time in seconds had elapsed before selecting a complete route between Diamond and Einstein.
Number of actions (NOA) refers to a count of all recorded clicks and can be interpreted as an indicator of overall activity. It does not, however, reveal anything about what the actions referred to.

Unique routes (UNIQ_R) refers to a count of how many unique routes between Diamond and Einstein a test-taker explored. It could be viewed as an indicator of effort towards solving the task. A route was defined as a simple path (paths that do not repeat nodes) between the nodes of Diamond and Einstein.

Repeated wrong routes (RWR) counts repeated completion of faulty routes between Diamond and Einstein. The variable gives information related to difficulties in keeping track of routes already tried and, together with unique routes, indicates efforts aimed towards solving the task.

Actions per minute (APM) was used as an indicator of the speed by which students executed actions. The variable was calculated starting from the first click on a road to reduce influence of any early exploratory, or accidental, clicks on some other place of the graphical interface. A lower APM might be related to more deliberation between the selection of actions and routes. (See Table 2 for descriptive statistics on the variables that were developed and used in clustering.)

Performance measure
Students’ mathematics score was chosen as performance variable since mathematics was the major domain for PISA 2012. Due to a matrix sampling design, not all students respond to the same items in the PISA test, and, hence, raw test scores are not reported. Instead, the performance scores in PISA are reported in the form of plausible values. Plausible values are imputed scores, estimated from student performance on PISA items and background information. In PISA 2012, each student has five plausible values, where each plausible value is a random draw from a distribution of “possible scores” for that particular student (OECD, 2014c). All plausible values were used in the analyses. They were standardised and centred to have a mean of 0 and a standard deviation of 1.

Self-reported effort
In PISA 2012, self-reported effort was assessed through an “effort thermometer”. Students were asked to rate, on a scale from 1 to 10, how much effort they spent on the PISA test, where 10 referred to maximum effort invested, see Figure 2.

Table 2. Descriptive statistics of the variables used in clustering.

| Variable                      | Median | Q₁    | Q₃    | Min | Max |
|-------------------------------|--------|-------|-------|-----|-----|
| Time on task (TOT)            | 68.40  | 44.70 | 106.80| 1.5 | 623.50 |
| Time to first action (T2FA)   | 14.60  | 10.70 | 21.30 | 0.0 | 534.20 |
| Time to first route (T2FR)    | 24.10  | 17.35 | 37.70 | 0.0 | 451.40 |
| Number of actions (NOA)       | 28     | 15    | 52    | 0   | 300  |
| Repeated wrong routes (RWR)   | 0      | 0     | 1     | 0   | 20   |
| Unique routes (UNIQ_R)        | 3      | 2     | 4     | 0   | 14   |
| Actions per minute (APM)      | 37.34  | 27.77 | 47.37 | 0.0 | 89.27 |

Note: Q₁ and Q₃ refer to the first and third quartile, respectively.
In brief, the analyses were carried out as follows. First, cluster analysis was used to find profiles suggesting different levels of effort (Research Aim 1). Then, Bayesian regression modelling was used to estimate if and how the mean mathematics performance varied by cluster (Research Aim 2), and if and how self-reported effort varied by cluster (Research Aim 3).

**Clustering method**

We chose to use k-means clustering since it has been successfully used in previous studies to find patterns of behaviour in computer-based assessments (Goldhammer et al., 2017; He, Liao, & Jiao, 2019; Qiao & Jiao, 2018), game-based assessments (Fossey, 2017; Polyak et al., 2017), and game-playing behaviours (Bauckhage et al., 2015, 2014; Drachen et al., 2012). The method of *k-means* is an *unsupervised* approach to data analysis, which aims to find groups of cases that are similar to each other with respect to some variables, without having any previously labelled data to guide the analysis (see Géron, 2019).

Two indicators were used to guide the selection of the number of clusters. The first indicator was the “elbow” method, which compares the total within sum of
squares (WSS) between different numbers of clusters (James et al., 2017). The second indicator was to examine the average silhouette widths for the different numbers of clusters. The average silhouette width is a measure ranging from −1 to 1. Silhouette scores close to 1 indicate that an instance lies close to its fellow cluster members and far away from members of the next closest cluster. Negative values indicate that the instance is closer to members of other clusters (Géron, 2019).

Before clustering, variables were transformed with Yeo-Johnson power transformation to improve normality of data (Yeo & Johnson, 2000), and were centred and standardised over all participants, since k-means clustering is sensitive to non-spherical data (Géron, 2019). K-means clustering was executed with the function named `kmeans` that comes with the built-in stats-package in R. It was parameterised with the “Hartigan-Wong” algorithm (Hartigan & Wong, 1979), and set to use 1,000 starting iterations to minimise the risk of finding suboptimal local minima. While results were stable with different seeds, a specific seed was set for the analysis to be reproducible.

The initial clusters were split with respect to the item score, since the interpretations of the level of effort would be different depending on whether a test-taker solved, or failed to solve, the task. If the task was solved, the level of effort was mainly interpreted as efficiency. If the task was not solved, the level of effort was mainly interpreted as level of test-taking motivation.

**Bayesian regression modelling**

When looking at relationships between clusters (defined by behavioural effort), test performance, and self-reported effort, a Bayesian approach to statistical data modelling was chosen. This approach allows for a probabilistic interpretation of results, makes analyses of multiple comparisons convenient, and provides a richer description of the uncertainty of estimations (for a brief introduction to Bayesian statistics and modelling, see van de Schoot et al., 2021).

**Test performance predicted by cluster (Model 1).** A hierarchical Bayesian approach was used to estimate how each of the clusters’ test performance depended on behavioural effort:

\[
y_i \sim \text{Normal}(\mu_i, \sigma_y)
\]

\[
\mu_i = \alpha + \sum_{c=1}^{8} \beta_c x_c(i)
\]

\[
\alpha \sim \text{Normal}(0, 1)
\]

\[
\beta_c \sim \text{Normal}(0, \sigma_b)
\]

\[
\sigma_b \sim \text{HalfNormal}(1)
\]

\[
\sigma_y \sim \text{HalfNormal}(1)
\]
where \( y \) denotes the test performance score, the subscript \( i \) denotes observations 1, \( \ldots, N \), and \( c \) refers to clusters. The performance scores (plausible values) were assumed to be normally distributed; the prior for the grand mean intercept \( \alpha \) was set to come from a normal distribution with a mean of 0 and a standard deviation of 1. The prior on the \( \beta \) parameters, which control the clusters deviation from the grand mean, were set to a normal distribution with a mean of 0 and a standard deviation \( \sigma_b \). A hyperprior on \( \sigma_b \) was a half-normal distribution with a standard deviation of 1; this prior can theoretically produce values that range from 0 to \( \infty \), but puts the 95% of probability on values between 0 and 1.96, and has a median of 0.67. We chose this prior since it reflects a range of differences that are in the plausible range with respect to previous empirical results reviewed in the introduction, while at the same time allowing for greater differences to be observed. The prior is a weakly informative prior (see https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations). As the PISA technical report (OECD, 2014c) recommends that analyses of plausible values be carried out on all plausible values, the model was first fitted with the first plausible value, and then updated with information from the remaining plausible values.

**Self-reported effort predicted by cluster (Model 2).** The self-reported test-taking effort was assumed to come from a continuous latent variable representing test-takers subjective test-taking effort, and was modelled using ordinal regression with a cumulative model. Ordinal regression was chosen since it can avoid associated errors using metric models on skewed ordinal data (Liddell & Kruschke, 2018). For more details about the model which we used to analyse the self-reported data, see Bürkner and Vuorre (2019).

The cumulative probit model assumes that the probability of responding to a specific category \( P(Y = K \in \{1, 2, \ldots, 10\}) \) comes from a latent, normally distributed variable \( \tilde{Y} \), with a mean of 0 and a standard deviation of 1, which has been partitioned by threshold parameters \( \tau_k \). The thresholds are modelled to be dependent on the coefficients \( b \) to create a regression model. In our case:

\[
\tilde{Y}_{ik} = \eta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \quad k = 1, 2, \ldots, 9
\]

\[
\eta_i = b_{11}x_{i1} + b_{12}x_{i2} + \ldots + b_{17}x_{i7}
\]

where the predictor variables \( x \) are dummy variables given by the clusters. The probability \( \tilde{Y}_{ik} \) of giving a response to a specific category can be retrieved by the cumulative normal distribution function, but, in this case, we are not primarily interested in the probabilities of selecting a particular response themselves, but how the latent distribution varies with the behavioural effort. Two variations of this model were fitted, one using low_effort_0 as contrast (Model 2a), and the other using low_effort_1 as contrast (Model 2b).
Posterior distributions for both Model 1 and Model 2 were estimated with Markov chain Monte Carlo (MCMC) using Hamiltonian Monte Carlo and NO-U-TURN Sampler, with four chains, 1,000 warmup samples, and a total of 8,000 iterations.

**Statistical software, code repository, and reproducibility**

All data cleaning, pre-processing, creation of variables, and data visualisation was done in the R environment (R Core Team, 2018). Bayesian regression modelling was run from the brms package (Bürkner, 2017), which uses the Stan programming language (Carpenter et al., 2017) to compute the posteriors. The ComplexHeatmap package (Gu et al., 2016) was used to create the heat map visualisations.

A repository containing the data and code necessary to recreate the analyses, figures, and tables, is available at: https://osf.io/hfsjy/ (Lundgren, 2021), to ensure scientific transparency and reproducibility of analyses (Nosek et al., 2015).

**Results**

**K-means clustering**

There was no clear single “kink” in the elbow plot, but the plot suggested 2 or 4 clusters (see Figure 3). The average silhouette widths suggested 4 clusters. A cluster size of 4 was chosen since it was compatible with both indicators and feasible to interpret with respect to different levels of effort.

**Clustering results**

The k-means clustering resulted in a total of 4 clusters. The clusters were labelled according to their most salient characteristics (described in Table 3), namely: high_effort, mid_effort, low_effort, and planner. Clusters were then split into pairs depending on the item score (correct/incorrect), indicated by a 1 or 0 at the end, effectively creating 8 clusters. Thus, high_effort_1 refers to the group displaying high effort and solving the task, and, similarly, low_effort_0 refers to the group displaying low effort and failing to solve the task.

Figure 3. Elbow plot in left panel (A), average silhouette widths in right panel (B).
The clusters varied in size, with the largest cluster mid_effort_1 (displaying medium effort and solving the task) constituting 39% of the total sample, and the smallest planner_0 at 3.4% of the total sample. Table 3 includes a description of salient characteristics of the different clusters and how these characteristics were interpreted in terms of effort, efficiency, and motivation.

The characteristics of the clusters were derived from interpreting the clusters’ variable values (see Table 4, where descriptive statistics in the form of unstandardised median values are displayed), as well as inspecting the heat map in Figure 4, where the behaviour of each individual student in each cluster is visualised in terms of their activities within the item. Brighter fields suggest more activity (longer time on task, larger number of actions, etc.), while darker fields suggest less activity. Most clusters could be characterised in terms of level of effort relative to the other clusters (see Table 3 and Figure 4). The planner clusters were, however, more difficult to interpret with respect to effort. The planners displayed a pattern of approaching the task where students had a long time of inactivity before starting to interact with the task, around 30 seconds before any action and around a minute to complete the first route, which often was the correct solution. This behaviour could be interpreted as if the student was planning their actions before executing them, and their efforts hence not easily observed.

| CLUSTER   | N   | % of sample | Interpretation of cluster characteristics |
|-----------|-----|-------------|------------------------------------------|
| high_effort_0 | 208 | 6.4         | Spent a long time on the task, performed many actions, tried many routes, and repeated wrong routes. This indicated **high effort** before giving up the task, and a **high motivational magnitude**. |
| high_effort_1 | 618 | 19.1        | Spent a long time on task, performed many actions, tried many routes, and repeated wrong routes, before solving the task. This indicated **high effort** and **lower efficiency** of problem-solving strategy. |
| low_effort_0 | 330 | 10.2        | Minimal activity on most variables, tried no or few routes before ending the task. This indicated **low effort** before giving up the task, and a **lower motivational magnitude**. Some of the test-takers might not have engaged in trying to solve the task at all. |
| low_effort_1 | 163 | 5.0         | Spent a short time on task, performed few actions, tried few routes, and repeated no or few wrong routes, before solving the task. This indicated **low effort** and a **higher efficiency** of problem-solving strategy. |
| mid_effort_0 | 150 | 4.6         | Spent an average amount of time, performed an average number of actions, and tried an average number of routes before giving up. This indicated an **average effort** before giving up the task, and an **average motivational magnitude**. |
| mid_effort_1 | 1265 | 39.2       | Spent an average amount of time on task, an average number of actions, and tried average amount of routes, solving the task. This indicated an **average effort**, and **average efficiency** of problem-solving strategy. |
| planner_0   | 109 | 3.4         | Spent a long time on a task, a long time to the first action, performed few actions and tried few routes before giving up on the task – as if they were first planning a solution before starting to interact. This indicated an **average or low effort**, and thus an **average or low motivational magnitude**. |
| planner_1   | 388 | 12.0        | Spent a long total time, a long time to the first action, performed few actions, and tried few routes before solving the task – as if they were first planning a solution and then delivered it. This indicated an **average** to **high effort**, and **average to low efficiency**. |
| CLUSTER      | TOT   | T2FA  | T2FR  | NOA   | RWR  | UNIQ_R | APM   |
|--------------|-------|-------|-------|-------|------|--------|-------|
| high_effort_0| 133 (97, 195) | 14 (10, 21) | 25 (19, 36) | 80 (50, 114) | 2 (1, 4) | 4 (3, 7) | 42 (34, 51) |
| high_effort_1| 113 (83, 153) | 14 (10, 19) | 23 (18, 32) | 70 (51, 97)  | 1 (1, 3) | 5 (4, 7) | 44 (36, 53) |
| low_effort_0 | 37 (24, 59)   | 13 (4, 19)   | 0 (0, 0)   | 7 (3, 13)   | 0 (0, 0) | 0 (0, 0) | 20 (9, 30)  |
| low_effort_1 | 32 (26, 37)   | 12 (10, 15)  | 23 (18, 29) | 7 (6, 10)   | 0 (0, 0) | 1 (1, 1) | 26 (22, 31) |
| mid_effort_0 | 56 (45, 75)   | 13 (10, 17)  | 23 (17, 33) | 27 (19, 45) | 0 (0, 0) | 2 (1, 3) | 40 (33, 48) |
| mid_effort_1 | 54 (41, 71)   | 14 (11, 18)  | 22 (18, 30) | 27 (18, 39) | 0 (0, 0) | 3 (2, 4) | 42 (34, 50) |
| planner_0    | 102 (72, 142)| 33 (20, 52)  | 66 (50, 117)| 20 (11, 33) | 0 (0, 0) | 1 (1, 2) | 24 (17, 28) |
| planner_1    | 100 (75, 150)| 40 (22, 62)  | 75 (54, 112)| 17 (11, 29) | 0 (0, 0) | 2 (1, 3) | 24 (16, 32) |

Note: Values displayed in seconds for the three columns to the left and as counts for the four columns to the right. TOT = time on task; T2FA = time to first action; T2FR = time to first route; NOA = number of actions; RWR = repeated wrong routes; UNIQ_R = unique routes; APM = actions per minute.
Descriptive statistics were further calculated for self-reported effort (assessed by the effort thermometer, see Figure 2), suggesting rather little variability across clusters. All clusters had a median of 8 (on the 10-point

Figure 4. Heat map of clustered standardised values (z).
Note: The clusters are on rows; variables are in columns. Within clusters, each coloured row represents the values of one student. The individual observations are coloured with respect to their z values as indicated by the colour of the bar (z) to the right. The discrete dark grey and light grey annotations in the rightmost column indicate correct and incorrect item scores, respectively. TOT = time on task; T2FA = time to first action; T2FR = time to first route; NOA = number of actions; RWR = repeated wrong routes; UNIQ_R = unique routes; APM = actions per minute.
effort thermometer), and the inter-quartile range (IQR) was between 7 and 9 for of all clusters, except low_effort_0, which had an IQR between 6 and 9. Previous studies have also noted that the “typical response” to the effort thermometer is an 8, and that the majority of responses tend to fall between 6 and 9 (Butler & Adams, 2007; Skolverket [Swedish National Agency for Education], 2015).

**Bayesian regression models**

**Test performance predicted by cluster (Model 1)**

The results from the regression show that the low_effort_0 cluster had the largest negative effect on the mean test score, in ascending order followed by planner_0, mid_effort_0, and high_effort_0. The clusters planner_1, high_effort_1, mid_effort_1, and low_effort_1, had positive effects on the estimated mean score (also in ascending order). Model results are displayed in Table 5.

Pairwise comparisons showed that the overall differences of mean achievement scores between clusters ranged from near 0 to over 1 standard deviations (SD). All clusters that got the item correct had higher overall performance scores compared to clusters that answered the item incorrectly. Comparing the clusters that did not solve the task (the “0” clusters), the point estimate of mean differences between medium effort and low effort was 0.41 with a 95% credible interval (CI) between [0.24, 0.58]. This can be interpreted as groups spending medium amounts of effort on the task, and failing to solve it is most likely to perform around 0.41 SD higher on the PISA math test compared to a group investing low amounts of effort before failing to solve the task, and the model implies a 0.95 probability that the difference lies between 0.24 and 0.58. The estimated difference between high effort and low effort was estimated to 0.63 (95% CI = [0.48, 0.79]). The estimated difference between high effort and medium effort was 0.22 (95% CI = [0.03, 0.41]). To better understand the extent of these

| Parameter          | Estimate | Est. Error | Lower CI | Upper CI |
|--------------------|----------|------------|----------|----------|
| $\alpha$           | -0.19    | 0.2        | -0.58    | 0.21     |
| $\beta_{\text{low\_effort}_0}$ | -0.12    | 0.21       | -0.54    | 0.28     |
| $\beta_{\text{high\_effort}_0}$ | 0.38     | 0.2        | -0.03    | 0.77     |
| $\beta_{\text{mid\_effort}_0}$ | -0.75    | 0.2        | -1.16    | -0.35    |
| $\beta_{\text{low\_effort}_1}$ | 0.53     | 0.21       | 0.12     | 0.95     |
| $\beta_{\text{mid\_effort}_1}$ | -0.34    | 0.21       | -0.77    | 0.07     |
| $\beta_{\text{high\_effort}_1}$ | 0.48     | 0.2        | 0.07     | 0.87     |
| $\beta_{\text{planner}_0}$     | -0.43    | 0.21       | -0.87    | -0.02    |
| $\beta_{\text{planner}_1}$     | 0.18     | 0.2        | -0.24    | 0.58     |
| $\sigma_y$         | 0.56     | 0.17       | 0.32     | 0.99     |
| $\sigma_0$         | 0.91     | 0.01       | 0.89     | 0.93     |

Note: Est. Error refers to the standard deviation of the posterior. Lower and upper credible interval (CI) indicate 95% CI based on the 2.5% and 97.5% percentile, respectively. The $\alpha$ coefficient refers to the estimated intercept in the model (the grand mean). The value of the $\beta$ coefficients indicates the effect each cluster has on test performance. The $\sigma_y$ parameter is the estimated standard deviation of the observed scores. The $\sigma_0$ parameter can be interpreted to represent the between-group variability in score.
differences, it can be noted that a difference of 1 SD would indicate a performance difference of around 100 points on the original PISA scale. PISA also reports results through categorisation of more qualitative proficiency levels (six levels, with two sub-levels on the lowest level suggesting only very basic knowledge; OECD, 2014b), and one step on the PISA 2012 mathematics proficiency level scale corresponds to roughly 60 points, which is close to the estimated difference in mean score between high_effort_0 and low_effort_0. Thus, among the clusters that failed to solve the task, their estimated level of behavioural effort on the task was clearly related to overall test performance: The more effort they spent on the traffic task, even if they did not solve it, the better their overall performance.

Among the clusters that solved the task (the “1” clusters), the estimated mean differences between medium effort and low effort was −0.06 (95% CI = [−0.21, 0.09]), indicating a small negative relationship level between effort and test performance, but with a greater uncertainty indicated by the 95% interval including 0. The difference between high effort and low effort was estimated at −0.16 (95% CI = [−0.32, 0.00]). The estimated mean difference between high effort and medium effort was −0.10 (95% CI = [−0.18, −0.01]). Thus, among the clusters that solved the task, there were weak negative relationships between behavioural effort on the task and overall performance. Thus, the group of students who solved the task with little effort performed slightly better overall. Estimates were, however, more uncertain for these clusters (see Figure 5 for a visualisation of the estimated differences).

**Self-reported effort predicted by cluster (Model 2)**

To investigate whether behavioural effort was related to self-reported effort, an ordinal regression was run with cluster as independent variable and latent self-

![Figure 5](image-url)

**Figure 5.** Pair-wise comparisons of posterior mean test performance with respect to level of effort and item score.

Note: Error bars show a 95% credible interval (CI). The dark grey distribution included a difference of 0 within the 95% interval, indicating a considerable uncertainty with respect to the direction of the difference.
reported effort as dependent variable. Figure 6 displays estimated $\beta$ coefficients compared to the contrasting cluster (see Appendix 1 for numerical representations of results). In this model, the $\beta$ coefficients can be interpreted as the effect the cluster had on the subjective rating of test-taking effort compared to the contrasting cluster.

From inspecting the left plot in Figure 6, which shows results from Model 2a, we can see that all clusters had a positive effect on latent test-taking effort compared to low_effort_0. Hence, all clusters were estimated to have a higher self-reported latent effort than the low-effort cluster that did not solve the task. The largest effect was observed for low_effort_1 (the cluster that solved the task with little effort) with $\beta = 0.39$ (95% CI = [0.19, 0.58]), and the smallest effect was for planner_0, $\beta = 0.24$ (95% CI = [0.01, 0.47]).

In the right plot in Figure 6, which shows results from Model 2b, it is clearly visible that the distribution of low_effort_0 is more offset from the contrasting cluster low_effort_1 compared to the other clusters. The results indicate that the cluster that, according to within-item behaviour, spent little effort on the Traffic task and gave up without solving it, low_effort_0, also had the lowest level of self-reported effort, while the other clusters did not have a clearly visible relationship with self-reported latent effort.

**Discussion**

The aim of the present study was threefold. First, our ambition was to extract variables from raw log-file data describing test-takers behaviours on a PISA 2012 problem-solving item, cluster the variables, and, if possible, interpret the resulting clusters in terms of level of effort. Second, the study aimed to
investigate how test performance varied as a function of cluster. The third aim was to investigate how self-reported test-taking effort varied as a function of cluster. In brief, the results indicate that there is a lot of variation with respect to students’ ways of responding to the traffic item, but that the students’ behaviours could indeed, after extraction of variables and clustering, be interpreted as indicative of different levels of effort put into trying to solve the task. A common assumption, and an assumption that has received ample empirical support in previous studies, is that a higher level of effort is related to a higher level of performance. In the present study, overall performance (here assessed through students’ performance in mathematics in PISA) was, however, not in a straightforward way related to amount of behavioural effort displayed on the item. Rather, it seems as if a more nuanced interpretation of test-taking effort is warranted if effort is to be interpreted as an indicator of test-taking motivation. In this study, the low-effort cluster that solved the task (i.e., the most efficient cluster) was the highest performing cluster. However, the low-effort cluster that failed to solve the task (i.e., gave up the task without using much resources) was the lowest performing cluster, and also the cluster with the lowest self-reported latent effort. Below, findings are discussed with respect to possible reasons underlying the observed differences in effort, as well as the clusters’ relation to performance and self-report measures. We also discuss the contribution of considering both time and action variables derived from log-file data when assessing test-taking effort.

Did the clustering of process data reveal anything about effort?

The k-means clustering of variables derived from the logged behavioural traces resulted in a total of four clusters. Three of the clusters were interpreted as showing either low, medium, or high effort on the traffic task. The low-effort cluster was interpreted as such because students in this cluster had used fewer resources relative to the other clusters. The low-effort cluster displayed short total times, fewer numbers of actions, and fewer numbers of unique routes tried, compared to the medium-effort cluster, and the high-effort cluster used even more resources. There was also a cluster with a different kind of response pattern, labelled “planner”. Test-takers in this cluster spent comparably longer time before starting to execute actions, acting as if they were planning a solution before they delivered it. This different pattern of behaviour made the planner cluster difficult to interpret in terms of effort. The design of the traffic item allows for “planner” behaviour to occur, since it does not necessitate any behavioural interaction (that is recorded in the log-file) in the search for a correct solution. Test-takers can spend their time inspecting the map visually and conduct mental arithmetic and comparisons of paths, perhaps taking notes. Only when they have already identified the solution will they deliver it by clicking, which are the only behaviours that are recorded in the
log file. The planner cluster clearly illustrates a key difficulty with respect to the inferential utility of response process data derived from an item designed such that it can be solved without leaving much of a behavioural trace. However, apart from the planners, the results indicated that it is possible to retrieve profiles describing different levels of within-item effort by using process data from human–computer interaction.

**How did behavioural effort relate to PISA performance?**

As we hypothesised that effort spent on the task could indicate either efficiency or test-taking motivation depending on whether the test-taker solved the task or not, all clusters were split into pairs based on item score (correct/incorrect). Comparisons between the clusters that did not solve the task showed that higher levels of behavioural effort were associated with higher estimated mean test performances. Students that were more perseverant, and failed this task after having tried hard, did better overall. Differences were in some comparisons substantial, and they are in the same range as the effect sizes reported by Wise and DeMars (2005; see also Silm et al., 2020). These results are consistent with the interpretation that the level of effort invested before giving up the task is an indication of the level of test-taking motivation.

Among the clusters that solved the task, a negative difference in mean performance was found between the high-effort and medium-effort cluster, as well as between the high- and low-effort cluster, which means that a lower task effort was associated with a higher test performance. These results are consistent with the interpretation that the effort they actually needed to invest to solve the task may not say very much about their motivation. We simply do not know how much more effort they would have been willing to spend, only that they spent as much effort as they needed to solve the item. The differences in effort and efficiency could be a function of skill, or perhaps luck, in finding the correct route.

So, overall, the results from regressing test performance on behavioural within-item effort suggest that the level of effort spent before giving up tasks can tell us something about test-taking motivation, while the effort spent before presenting a correct answer may say more about efficiency.

**How did behavioural effort relate to self-reported effort?**

Results indicated that all clusters had a higher level of self-reported latent test-taking effort compared to a cluster that displayed low effort and did not solve the task (low_effort_0). Students in this cluster did not only behave in a way that suggested low effort, they also self-reported spending less effort than students in other clusters. The differences derived from self-reported values support the interpretation that the low behavioural effort among test-takers that failed to solve the item indicates relatively lower test-taking motivation.
For many of the other comparisons in the regression analyses, there were no clear differences between clusters in terms of self-reported effort. This might suggest that self-reported test-taking effort measures different aspects of effort compared to behavioural effort (Silm et al., 2020). This could be especially true in the present case, where behavioural effort is observed from a single item, whilst the subjective effort concerns the full test. It should also be noted that the effort thermometer is a rather rough measure of self-reported effort with less variability than a composite effort scale may have (see Butler & Adams, 2007; Skolverket, 2015). Previous studies comparing relationships between self-reported and behavioural effort have found small to moderate correlations (Silm et al., 2020), though all previous studies have used other self-report measures and other ways of estimating behavioural effort compared to the present study. Even if our findings are somewhat inconclusive regarding convergence between self-reported and behavioural measures of effort, results clearly showed that the cluster that was interpreted as “the least motivated” cluster from analysis of task behaviour (low_effort_0) was also the cluster with the lowest self-reported effort. Although this is weak support for convergent validity, it does call for further studies of the convergence between different measures of self-reported versus observed test-taking effort.

Towards an improved understanding of the relationship between effort and motivation

As described in the introduction, a common way of objectively assessing test-taking effort is through response times, where a response time below a threshold is judged to be indicative of disengagement, non-solution behaviour, and low effort. Although response time is a well-established unobtrusive measure of effort, in certain assessment contexts and for certain item types, a fuller account of student within-item behaviour may allow for more valid conclusions regarding effort and motivation. Preliminary analyses on the current data suggest that, if the present study had applied a response time analysis with a liberal cut-off at 10 seconds, only a few (n = 30) cases of non-solution behaviours would have been observed. All cases would have come from the low-effort cluster that did not solve the task, but would only represent about 9% of all cases in this cluster. However, it cannot be decided at this point which of the classifications would be more correct. It is possible that classifications based on a response time measure compared to a more comprehensive process data measure might be underestimating the number of “non-effortful” responses for certain item types. More research on this issue is needed, but in some assessment contexts, it could be worth considering incorporating information from within-item actions in addition to response time indicators, to more precisely identify responses likely to be driven by low test-taking motivation.
For example, in the present data, we observed a participant that performed many “valid” effortful actions (highlighting different path segments), but ended up selecting no route between Diamond and Einstein. When exploring the log file, it turned out that the test-taker had simply highlighted all of the edges in the road network. Even if this behaviour were atypical, it illustrates how effortful behaviours driven by other motivations than test-taking motivation could be mistaken for motivated behaviour. Thus, behavioural effort expended on test items is not an infallible indicator of test-taking motivation.

Limitations and future research

The present study only considered process data from one item in one particular context, and findings cannot be generalised to other items and contexts. Preferably, multiple items should be used in similar analyses to be able to properly generalise the overall test-taking effort. Future studies should try to accumulate evidence from multiple interactive items to improve the assessment of overall test-taking effort. The outcome measure in PISA, plausible values, may not be an optimal answer to how well students performed on the actual test, since it is estimated partly by imputation and the use of other informational sources, which makes it less reliable for individual students and small subgroups of students. The differences in behavioural effort and efficiency among test-takers who solved the item warrants an explanation; there are, perhaps, better or worse strategies in solving the task. To address these questions, computational cognitive modelling could be used to better understand the cognitive processes that might underlie differences in task performance (Kane & Mislevy, 2017; LaMar, 2014). Further studies could use even richer measures of behaviour, such as eye-tracking, as well as qualitative observations, and verbal reports. Measures like these could (a) further improve our understanding of test-taker behaviour and reasons for different behaviours and (b) provide convergent validity evidence with respect to effort assessed from log-file data; a similar conclusion is made by Goldhammer et al. (2014).

Conclusions

One key finding from the present study is that response process data from human–computer interactions can be used to shed further light on the complex relationship between test-taking effort and test-taking motivation. Exploring these data can yield both theoretical and methodological contributions. Another key finding is that within-item behaviours can be relevant to consider alongside other measures of test-taking effort such as response times and self-reports. The main key finding from this study is, however, that the level of effort that is diagnostic of low test-taking motivation seems to be low levels of effort invested before giving up a task, while low levels of effort
before solving a task may not be diagnostic of test-taking motivation. Just as the character of a Star Trek commander is tested in the Kobayashi Maru scenario – a no-win scenario – perhaps the motivational character of test-takers can best be revealed by items they cannot solve. Although managing lower level behavioural data introduces a fair share of complexities, it seems like a promising way to improve measurements of test-taking effort and motivation, as well as other constructs.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Data and code availability statement**

The data that support the findings and code to reproduce the results of this study are openly available at open science framework (OSF) at https://osf.io/hfsjy/, https://doi.org/10.17605/OSF.IO/HFSJY

**Notes on contributors**

**Erik Lundgren** is a doctoral student at the Department of Applied Educational Science, Umeå University, Sweden. He is a member of the research group Measurement in Behavioural Sciences. His doctoral research project involves using response process data to complement, enhance, and validate inferences from large-scale assessments.

**Hanna Eklöf** is an associate professor at the Department of Applied Educational Science, Umeå University, Sweden. She has a PhD in educational measurement from Umeå University. Her research interests are in the field of educational psychology and, more specifically, the psychology of test-taking, test-taking behaviour and motivation, large-scale testing, and assessment quality issues. She is currently the principal investigator of a project funded by the Swedish Research Council, where student effort and test-taking behaviour is investigated through self-report and process data in PISA. She is also co-investigator in projects investigating math anxiety, working memory, and performance.

**ORCID**

Erik Lundgren  http://orcid.org/0000-0002-6998-3397
Hanna Eklöf  http://orcid.org/0000-0002-4630-6123

**References**

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Atkinson, J. W. (1964). An introduction to motivation. Van Nostrand.
Bauckhage, C., Drachen, A., & Sifa, R. (2015). Clustering game behavior data. *IEEE Transactions on Computational Intelligence and AI in Games, 7*(3), 266–278. https://doi.org/10.1109/TCIAIG.2014.2376982

Bauckhage, C., Sifa, R., Drachen, A., Thurau, C., & Hadiji, F. (2014). Beyond heatmaps: Spatio-temporal clustering using behavior-based partitioning of game levels. In *2014 IEEE Conference on Computational Intelligence and Games* (pp. 1–8). IEEE. https://doi.org/10.1109/CIG.2014.6932865

Buckley, B. C., Gobert, J. D., Horwitz, P., & O’Dwyer, L. M. (2010). Looking inside the black box: Assessing model-based learning and inquiry in BioLogica™. *International Journal of Learning Technology, 5*(2), 166–190. https://doi.org/10.1504/IJLT.2010.034548

Bürkner, P.-C. (2017). *brms: An R package for Bayesian multilevel models using Stan*. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science, 2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement, 8*(3), 279–304.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Cocea, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-learning environments. *User Modeling and User-Adapted Interaction, 19*(4), 341–385. https://doi.org/10.1007/s11257-009-9065-5

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

Drachen, A., Sifa, R., Bauckhage, C., & Thurau, C. (2012). Guns, swords and data: Clustering of player behavior in computer games in the wild. In *2012 IEEE Conference on Computational Intelligence and Games (CIG)* (pp. 163–170). IEEE. https://doi.org/10.1109/CIG.2012.6374152

Eklöf, H., & Knekta, E. (2017). Using large-scale educational data to test motivation theories: A synthesis of findings from Swedish studies on test-taking motivation. *International Journal of Quantitative Research in Education, 4*(1–2), 52–71. https://doi.org/10.1080/10508406.2017.086499

Fossey, W. A. (2017). *An evaluation of clustering algorithms for modeling game-based assessment work processes* [Doctoral dissertation, University of Maryland]. https://doi.org/10.13016/M2599Z34K

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O’Reilly.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist, 50*(1), 43–57. https://doi.org/10.1080/00461520.2014.999919

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students’ science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education* (pp. 407–425). Springer. https://doi.org/10.1007/978-3-319-50030-0_24
Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students’ minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1), 100–108. https://doi.org/10.2307/2346830

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 Education Working Paper No. 205). OECD Publishing. https://doi.org/10.1787/650918f2-en

He, Q., Liao, D., & Jiao, H. (2019). Clustering behavioral patterns using process data in PIAAC problem-solving items. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 189–212). Springer. https://doi.org/10.1007/978-3-030-18480-3_10

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R* (Corrected at 8th printing). Springer.

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 11–24). Routledge.

Karumbaiah, S., Baker, R. S., & Shute, V. (2018). Predicting quitting in students playing a learning game. In *Proceedings of the 11th Conference on Educational Data Mining* (pp. 167–176). International Educational Data Mining Society. https://eric.ed.gov/?id=ED593115

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619. https://doi.org/10.1177/0013164406294779

LaMar, M. (2014). *Models for understanding student thinking using data from complex computerized science tasks* [Doctoral dissertation University of California, Berkeley]. https://escholarship.org/uc/item/6d30w6th

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

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

Lundgren, E. (2021). *Within-item response processes as indicators of test-taking effort and motivation*. https://www.doi.org/10.17605/OSF.IO/HFSJY

Mills, C., Bosch, N., Graesser, A., & D’Mello, S. (2014). To quit or not to quit: Predicting future behavioral disengagement from reading patterns. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent tutoring systems* (Vol. 8474, pp. 19–28). Springer. https://doi.org/10.1007/978-3-319-07221-0_3
Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Organisation for Economic Co-operation and Development. (2014a). *PISA 2012 results: Creative problem solving: students’ skills in tackling real-life problems (Volume V)*. https://doi.org/10.1787/9789264208070-en

Organisation for Economic Co-operation and Development. (2014b). *PISA 2012 results: What students know and can do (Volume I, Revised edition, February 2014): Student performance in mathematics, reading and science*. OECD. https://doi.org/10.1787/9789264208780-en

Organisation for Economic Co-operation and Development. (2014c). *PISA 2012 technical report*.

Polyak, S. T., von Davier, A. A., & Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Frontiers in Psychology*, 8, 2029. https://doi.org/10.3389/fpsyg.2017.02029

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

Rafferty, A. N., Jansen, R. A., & Griffiths, T. L. (2020). Assessing mathematics misunderstandings via Bayesian inverse planning. *Cognitive Science*, 44(10), Article e12900. https://doi.org/10.1111/cogs.12900

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, 17(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Schwarz, N. (2007). Retrospective and concurrent self-reports: The rationale for real-time data capture. In A. A. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 11–26). Oxford University Press.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, 63, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Silm, G., Must, O., & Täht, K. (2019). Predicting performance in a low-stakes test using self-reported and time-based measures of effort. *TRAMES: A Journal of the Humanities & Social Sciences*, 23(3), 353–376. https://doi.org/10.3176/tr.2019.3.06

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, 31, Article 100335. https://doi.org/10.1016/j.edurev.2020.100335

Skolverket. (2015). *To respond or not to respond. The motivation of Swedish students in taking the PISA test*. https://www.skolverket.se/publikationer?id=3545

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Mårtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1, Article 1. https://doi.org/10.1038/s43586-020-00001-2

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1
Appendix 1. Summary of estimated parameters in Model 2a and Model 2b

| Parameter | Model 2a | | | Model 2b | | |
|-----------|----------|--|---|----------|--|---|
|           | Est.     | Err | Low | Upr | Est.     | Err | Low | Upr |
| $\tau_1$  | -2.17    | 0.09 | -2.36 | -1.99 | -2.56    | 0.11 | -2.78 | -2.35 |
| $\tau_2$  | -1.81    | 0.08 | -1.96 | -1.66 | -2.2     | 0.1  | -2.39 | -2.01 |
| $\tau_3$  | -1.52    | 0.07 | -1.65 | -1.38 | -1.9     | 0.09 | -2.08 | -1.73 |
| $\tau_4$  | -1.25    | 0.07 | -1.38 | -1.12 | -1.63    | 0.09 | -1.8  | -1.46 |
| $\tau_5$  | -0.9     | 0.06 | -1.03 | -0.78 | -1.29    | 0.08 | -1.46 | -1.12 |
| $\tau_6$  | -0.6     | 0.06 | -0.72 | -0.48 | -0.99    | 0.08 | -1.15 | -0.82 |
| $\tau_7$  | -0.06    | 0.06 | -0.18 | 0.07  | -0.44    | 0.08 | -0.6  | -0.28 |
| $\tau_8$  | 0.66     | 0.06 | 0.54  | 0.79  | 0.28     | 0.08 | 0.12  | 0.44 |
| $\tau_9$  | 1.5      | 0.07 | 1.37  | 1.63  | 1.12     | 0.08 | 0.95  | 1.28 |
| $\beta_{high\_effort\_0}$ | 0.23 | 0.1 | 0.04 | 0.42 | -0.39 | 0.1 | 0.58 | -0.19 |
| $\beta_{high\_effort\_1}$ | 0.33 | 0.07 | 0.19 | 0.48 | -0.15 | 0.11 | 0.37 | 0.06 |
| $\beta_{low\_effort\_0}$ | 0.39 | 0.1 | 0.19 | 0.58 | -0.05 | 0.09 | 0.23 | 0.13 |
| $\beta_{low\_effort\_1}$ | 0.26 | 0.11 | 0.05 | 0.47 | -0.13 | 0.12 | 0.36 | 0.11 |
| $\beta_{mid\_effort\_0}$ | 0.4 | 0.07 | 0.26 | 0.53 | 0.01 | 0.09 | -0.16 | 0.18 |
| $\beta_{mid\_effort\_1}$ | 0.24 | 0.12 | 0.01 | 0.47 | -0.15 | 0.13 | 0.13 | 0.11 |
| $\beta_{planner\_0}$ | 0.29 | 0.08 | 0.14 | 0.45 | -0.09 | 0.1 | -0.28 | 0.1 |
| $\beta_{planner\_1}$ | 0.29 | 0.08 | 0.14 | 0.45 | -0.09 | 0.1 | -0.28 | 0.1 |

Note: Posterior estimates of $\tau$ parameters and $\beta$ parameters for each cluster. The contrasting cluster is $low\_effort\_0$ for Model 2a, and $low\_effort\_1$ for Model 2b. The credible intervals (CI) are at 95%. Low = lower CI; Upr = upper CI; Est. Err = estimated standard deviation.