Large-scale Assessments in Education

RESEARCH Open Access

Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment

Yang Jiang*, Tao Gong, Luis E. Saldivia, Gabrielle Cayton-Hodges and Christopher Agard

*Correspondence:
yjiang002@ets.org
Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA

Abstract

In 2017, the mathematics assessments that are part of the National Assessment of Educational Progress (NAEP) program underwent a transformation shifting the administration from paper-and-pencil formats to digitally-based assessments (DBA). This shift introduced new interactive item types that bring rich process data and tremendous opportunities to study the cognitive and behavioral processes that underlie test-takers' performances in ways that are not otherwise possible with the response data alone. In this exploratory study, we investigated the problem-solving processes and strategies applied by the nation's fourth and eighth graders by analyzing the process data collected during their interactions with two technology-enhanced drag-and-drop items (one item for each grade) included in the first digital operational administration of the NAEP's mathematics assessments. Results from this research revealed how test-takers who achieved different levels of accuracy on the items engaged in various cognitive and metacognitive processes (e.g., in terms of their time allocation, answer change behaviors, and problem-solving strategies), providing insights into the common mathematical misconceptions that fourth- and eighth-grade students held and the steps where they may have struggled during their solution process. Implications of the findings for educational assessment design and limitations of this research are also discussed.

Keywords: Process data, Large-scale assessments, Problem solving, Problem-solving strategy, Mathematics, Cognitive and metacognitive processes

Introduction

Recent years have seen a rapid advancement in the use of technology and computers for mathematics learning and classrooms (Hoyles and Noss 2003; Koedinger and Corbett 2006). To address the rapid development and increasing importance of educational technology, large-scale assessments have started to transition from the traditional paper-and-pencil assessments to digitally-based assessments (DBA) (He, Borgonovi, and Paccagnella 2019; Scalise and Gifford 2006; Zenisky and Sireci 2002). For example, the National Assessment of Educational Progress (NAEP) started to

Springer Open

administer its mathematics assessments on handheld tablets in 2017. This transition offers tremendous opportunities for innovation through the introduction of new types of interactive and technology-enhanced items and mathematical tools as described in Table 1. Additionally, the digital testing platform used in NAEP DBA makes students' assessment experience significantly different from that in paper-and-pencil administrations. For example, the platform includes a tool bar that allows students to use tools such as a digital scratchpad, an on-screen calculator, and an equation editor to enter symbols, mathematical expressions and equations as part of their constructed responses. The interface also contains navigation icons to enable test-takers to move between items and zooming, theming, and text-to-speech features for testing accessibility.

The transition to digitally-based assessments and introduction of new item types also create opportunities for collecting rich process data (such as the detailed records of user interactions with the digital system and the timestamps of these user- or server-generated events) that are not available in traditional paper-and-pencil assessments. Process data produced from large-scale educational assessments afford the opportunity to study test-takers' paths to a solution and infer the cognitive and metacognitive processes they engage in at a fine-grained level (especially when they are combined with response data and theoretical frameworks (Mislevy, Almond, and Lukas 2003)), which the responses alone could not reveal (Provasnik 2021).

Table 1 Technology-enhanced items and mathematical tools that were first used in the 2017 NAEP digitally-based operational mathematics assessments for Grades 4 and 8

Item Type/Tool	Description
Multiple-Selection Multiple-Choice	This item type allows students to respond by selecting two or more choices that meet the condition stated in the stem of the item
Matching (Drag and Drop)	This item type allows students to respond by inserting (dragging and dropping) one or more source element(s) into target fields
Zones	This item type allows students to respond by selecting one or more region(s) on a graphic stimulus
Grid	This item type allows students to evaluate mathematical statements or expressions with respect to certain properties. The answer is entered by selecting cells in a table in which rows typically correspond to the statement and columns to the properties checked
Inline Choice	This item type allows students to respond by selecting one option from one or more drop-down menu(s) that might appear in various sections of an item
Interactive Ruler	This tool allows students to use an on-screen ruler to measure lengths of virtual objects on the screen to answer a question
Digital Calculator[a]	This tool allows students to use an on-screen calculator to perform operations needed to answer a question
Box and Whiskers	This tool allows students to create or modify a graphical five-number summary (box plot) of a numerical data set
Digital Scratchpad	This tool allows students to use their fingers or a stylus to perform computations, write notes, create hand drawings, annotate figures, highlight portions of a question, etc. on the touch-screen tablets
Equation Editor	This tool allows students to respond by entering numbers and mathematical expressions or equations using an onscreen pallet. A customized version of the equation editor is provided at each of grades 4, 8, and 12

[a] A digital four-function calculator was available for a selected set (approximately 1/3) of the items within one administration at Grade 4. A digital scientific calculator was available for a selected set (approximately 1/3) of the items within one administration at Grade 8 and Grade 12

For example, drag-and-drop (D&D) items have been increasingly used by test developers in digitally-based educational assessments (Arslan et al. 2020; Bryant 2017; Scalise and Gifford 2006). On D&D items, test-takers give a response by selecting and dragging sources into corresponding targets (see Figs. 2 and 3 for examples). Compared to the conventional multiple choice (MC) items, D&D items have been used to reduce the effect of random guessing, strengthen measurement, and improve test-taker engagement and motivation, considering its potential to better represent construct-relevant skills related to matching, categorizing, (re)ordering/(re)arranging, and sequencing (Arslan et al. 2020; Bryant 2017; Scalise and Gifford 2006). Process data on D&D items are rich and include the detailed records of student interactions with the system, such as their response actions and the timestamps of these actions.

One of the important goals for K-12 mathematics education is to help students develop knowledge and skills needed for mathematical problem solving (National Council of Teachers of Mathematics 2000). There is extensive evidence that individuals who apply efficient problem-solving strategies are more likely to be successful in academic performance and learning tasks (Pape and Wang 2003; Schoenfeld 1992). Research has also shown the effectiveness of providing instruction and/or feedback on problem-solving strategy in improving learners' ability to solve problems and their academic success, especially for those with low prior knowledge (Fyfe et al. 2012; Verschaffel et al. 1999). The examination of problem-solving strategies has been incorporated by researchers in mathematics curriculum to evaluate students' competency and to understand individual differences in mathematical problem solving (Cai et al. 2014). Similarly, research has documented the effect of instructions on metacognitive strategy on students' performance on solving mathematical problems (Kramarski et al. 2002; Özsoy and Ataman 2009). Therefore, it is crucial to understand and assess the cognitive and metacognitive processes involved in solving mathematical problems (Montague and Bos 1990) in educational assessments and identify students who struggle in these processes for further instruction and scaffolding.

Polya (1957) proposed that problem solving involves four phases: understanding the problem, devising a plan, trying and carrying out the plan, and monitoring and reflecting on the solution. During the past decades, researchers have extended Polya's four-phase framework and developed new models that are its variations to understand the cognitive and metacognitive aspects that underlie solution processes (Lester 1994; e.g., Schoenfeld 1992; Yimer and Ellerton 2010). In these frameworks, both cognition and metacognition were considered as integral to mathematical problem solving. Poor metacognitive skills such as not being able to monitor and regulate one's own solution process are obstacles to problem solving success even for students with rich knowledge in the content area (Goos 2002). The steps involved in mathematical problem solving also correspond to the model of self-regulated learning (SRL) that Winne and Hadwin (1998) developed. In this model, SRL is comprised of cyclical phases where students develop an understanding of the task, set goals and construct plans to achieve their goals, execute various tactics and strategies, metacognitively monitor and reflect on their learning process, and adapt their plans, behaviors, and strategies accordingly. Meanwhile, the use of strategies such as guess-and-check, visualization, and strategically utilizing tools like a calculator is deemed as crucial to solving mathematical problems.

Traditional measures of these processes and strategies involved in solving mathematical problems are mainly obtained through think-aloud protocols, structured interviews and observations. For example, Yimer and Ellerton (2010) interviewed 17 pre-service teachers as they engaged in solving mathematical problems. Based on the task-based interviews, they identified five phases of problem solving and the cognitive and metacognitive behaviors corresponding to each of the phases. In these phases, students make sense of a problem, transform the initial understanding into formulations of plans, implement the plans and explorations, evaluate the appropriateness of plans, actions, and solutions, and reflect on the solution process. In another study, Cai and Cifarelli (2005) examined the solution processes through videotaped protocols and self-reported measures of two college students when they worked on computer-based mathematics tasks. Despite their effectiveness, these traditional methodologies and measures are difficult to scale up and might not necessarily reflect authentic ways of mathematical problem solving. Process data collected from large-scale assessments, on the other hand, provide fine-grained information about how students plan, select, and execute various problem-solving strategies to find a solution and how they monitor and reflect on their response in an unobtrusive and scalable manner.

Despite the opportunities that process data from DBA afford beyond merely the student responses, process data have been mainly treated as a byproduct in educational assessments. There are relatively limited studies that examine how various cognitive and metacognitive processes and strategies manifest in large-scale mathematics assessments using process data. Bergner and von Davier (2019) reviewed a list of studies that analyzed NAEP process data collected from assessments predating its official transition to DBA in 2017 and proposed a five-level framework to describe the uses of process data. In this framework, process data use was ordered into five levels based on its relative importance in relation to outcome data alone, as shown in Fig. 1.

However, most of the relatively limited studies on process data from educational assessments either focus on the traditional MC items by analyzing answer change behaviors (Liu et al. 2015) or response time (Lee and Jia 2014), or involve action sequence analysis in more complex simulation-based science or engineering tasks (Gong et al. 2020; Han et al. 2019; Hao et al. 2015; Zhu et al. 2016). To our knowledge, no published study has examined the problem-solving processes and strategies on drag-and-drop items in large-scale mathematics assessments, despite its increasing use in DBA as a technology-enhanced item type.

An important piece of process data generated from drag-and-drop items is the sequences of test-takers' response actions (e.g., which source did a test-taker drag first and which target was the selected source dropped into, etc.). The response action

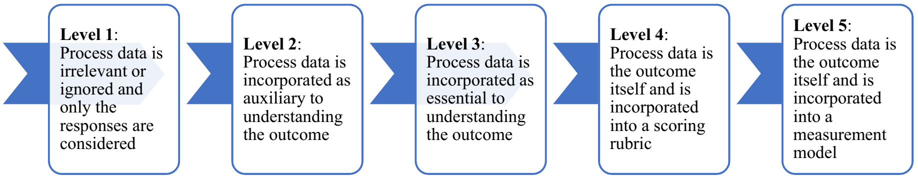
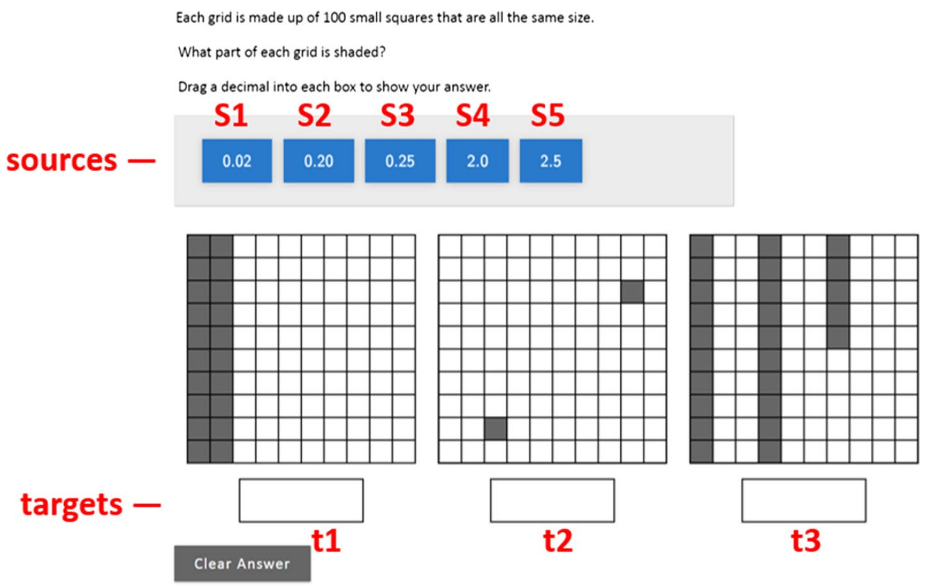

Level 1: Process data is irrelevant or ignored and only the responses are considered
Level 2: Process data is incorporated as auxiliary to understanding the outcome
Level 3: Process data is incorporated as essential to understanding the outcome
Level 4: Process data is the outcome itself and is incorporated into a scoring rubric
Level 5: Process data is the outcome itself and is incorporated into a measurement model

Fig. 1 Framework of process data uses in NAEP (Bergner and von Davier 2019)

sequences provide insights into the general and domain-specific strategies test-takers frequently plan and apply for solving problems. For example, a test-taker who was presented a D&D item as shown later in Fig. 2 (G4 item) might focus on the targets in the item and fill them sequentially. In other words, these students could start with the first target, conduct necessary mental computations, make a decision and drag a source to this target before they move on to focus on the second target and repeat the same procedure (i.e., fill target 1, target 2, and target 3 in order). We identify this approach as a *target-focused strategy*, in which the response sequence was organized by the visual representations in the targets, which were later transformed into symbolic representations and linked to the decimals in the sources. On the other hand, students could also focus on the sources and select and drag each source sequentially (i.e., *source-focused strategy*). For instance, they might start with the symbolic representation in the first source, perform computations, evaluate and make a decision on which target the source connects to, and execute the corresponding drag-and-drop action before moving on to the second source and repeat the same procedure. Other students would exhibit action sequences that do not show a systematic pattern of response behaviors. They might start with a source that was the easiest for them to solve, or could be randomly guessing or off-task. Students who submitted the same responses and thus received the same score on an item might adopt different strategies to generate a response, which are representative of the different underlying mental processes and possibly different levels of mathematical proficiency (e.g., understanding of the representations of numbers). This classification of response strategies that we developed was also adopted in our later work (Arslan et al. 2020) that was inspired by the current research, in which we examined the effect of drag-and-drop item design on student performance and strategy use.

This approach also enables us to study the efficiency of students' response strategies. For instance, the two-dimensional models in the targets on the G4 item are to-be-solved/



Each grid is made up of 100 small squares that are all the same size.

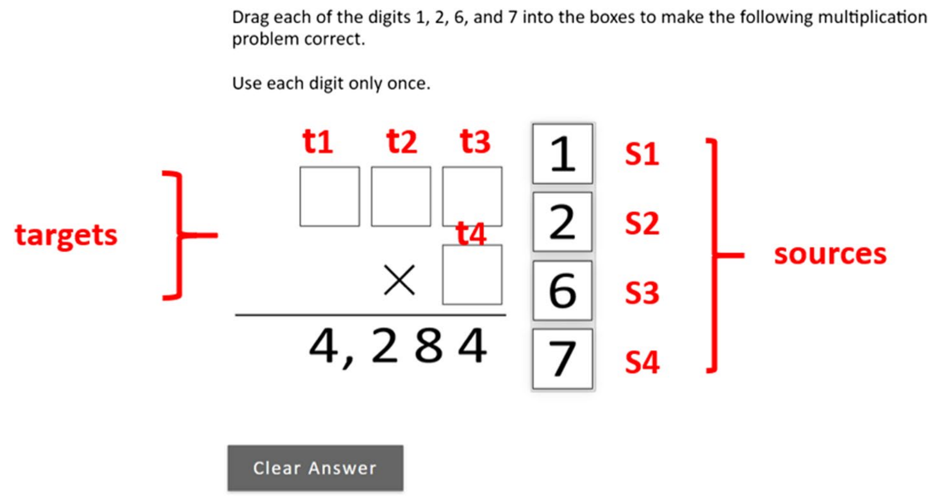
What part of each grid is shaded?

Drag a decimal into each box to show your answer.

Fig. 2 Screenshot of a released drag-and-drop item administered to fourth-grade students (G4 item) in the 2017 NAEP mathematics assessment. The sources and targets are denoted in red

converted mathematical objects while the sources are the symbolic representations to be matched. Applying a target-focused strategy is more efficient than a source-focused strategy in this case because once a target is translated into a decimal and matched with a source, test-takers do not need to perform the mental computations and decisions about the same target again, thus reducing their cognitive load (Sweller 1994). This approach requires three cognitive steps for students with high proficiency (mentally converting the two-dimensional model in the first target into a decimal and immediately filling this target with a corresponding source, then moving on to translate the second target, and repeating the same procedure for the third target). On the contrary, in a source-focused strategy, mental computations about the targets might need to be made more than once, whenever students evaluate a new source and compare it with the targets. Therefore, this strategy requires more cognitive steps even for students with high proficiency and is less efficient. Expert problem solvers typically search their strategy repertoire, evaluate the efficiency of possible strategies, and apply a strategy that is efficient and will aid in problem solving (Chi et al. 1982; Yimer and Ellerton 2010).

Similarly, action sequences displayed by students who responded to an item designed to evaluate eighth-grade students' problem-solving skills as shown in Fig. 3 (G8 item) would reveal the procedures and strategies used to solve the problem. For example, they could shed light on whether students solved the problem methodically and whether they used problem-solving strategies such as trial-and-error and guess-and-check. Trial-and-error is a strategy that is commonly used in mathematics practice (Elia et al. 2009). In this approach, students form a potential answer (could be either a complete or an incomplete response) and probably execute the relevant D&D actions, test it using mental calculation or mathematical tools, compare the results with the intended results, and repeat this procedure with another solution until the outcome of the computation matches the given product. Therefore, students who adopt this strategy would execute sequences that are longer than the minimum number of response actions required on this item (four D&D actions). Action sequences also reveal students' reasoning process. For instance,



Drag each of the digits 1, 2, 6, and 7 into the boxes to make the following multiplication problem correct.

Use each digit only once.

Fig. 3 Screenshot of a released drag-and-drop item administered to eighth-grade students (G8 item) in the 2017 NAEP mathematics assessment. The sources and targets are denoted in red

one potential reasoning process on the G8 item would be using the inverse operation of multiplication—division—to assist with problem solving. In this approach, the number to be placed in t4 has to be such that a three-digit factor is obtained when 4284 is divided by this number. Students who follow this strategy might place a source into t4 as their first step. Action sequence analysis helps us identify students who provided a correct answer but performed many unnecessary steps and adopted a less efficient strategy, considering the potential existence of partial knowledge that needs to be reinforced. Similarly, we could infer from process data where and when students who failed to provide a correct solution reached an impasse in attempting to solve the problem.

In addition to the sequence of actions executed by students, process data also record the timestamps of these actions. Numerous studies have explored total response time on MC items as an indication of motivation (Lee and Jia 2014; van der Linden 2008). On D&D items that involve a multi-step solution, process data enable us to further break the response completion process down into meaningful phases and explore the metacognitive processes and strategies involved in mathematical problem solving (Arslan et al. 2020; Gong et al. 2020). For example, students' first pause on an item (i.e., the time elapsed between entering an item and executing the first D&D action) is representative of the duration they spent on developing an understanding of the goal they need to achieve by reading the problem stem, setting a goal and constructing plans for problem solving, performing necessary computations and possibly executing strategies to solve the problem, and making a decision on the first D&D action (Arslan et al. 2020; Gong et al. 2020). Therefore, this measure might be related to the phases of defining task, and goal setting and planning in Winne and Hadwin's SRL framework (1998). On the other hand, a pause after finalizing one's answer until exiting an item (*last pause*) suggests that a test-taker might be metacognitively monitoring the formed response and reflecting on their solution process. In this sense, longer last pauses might indicate more time spent on monitoring and regulating one's behaviors and solution. The time between the first and the last drag-and-drop actions (*D&D execution time*) corresponds to the period when test-takers conducted additional necessary computations, possibly executed various strategies to solve the problem, and made decisions on the remaining response steps (Arslan et al. 2020; Gong et al. 2020). A long response execution time might indicate that the test-takers got stuck and reached an impasse, while a short execution time could be related to their high proficiency and efficiency, or caused by rapid guessing, speededness, carelessness, and disengagement (Guo et al. 2016; Lee and Jia 2014). Similar to the efficiency measures used in writing research (Galbraith and Baaijen 2019; Sandene et al. 2005), we calculated the average D&D execution time per response action to study the efficiency of the response process (Gong et al. 2020). Uncovering and identifying how students distribute their time in these phases will facilitate our understanding of their metacognitive competency and problem-solving processes for further intervention and instruction.

Current study

The current exploratory study analyzes process data from two mathematics drag-and-drop items in a large-scale educational assessment to investigate the following research questions.

RQ1: What cognitive and metacognitive strategies and processes do the nation's fourth- and eighth-grade students apply and engage in when solving drag-and-drop mathematics problems?

RQ2: Do students who received different scores also exhibit behaviors that are representative of different cognitive and metacognitive strategies and processes when solving the mathematics items? For example, do higher-scoring students adopt a more efficient problem-solving strategy?

To answer these research questions, we developed a list of process-based measures from the process data collected as fourth- and eighth-grade students interacted with two NAEP D&D items in the mathematics digitally-based assessment administered in 2017. These measures are representative of test-takers' cognitive and metacognitive processes and strategies during problem solving and include variables related to their response action sequences and time use. We aim to utilize these measures to infer how the fourth- and eighth-grade students in the United States responded to the technology-enhanced items, the misconceptions and struggles they had, and the problem-solving strategies they executed. These measures were later compared across the students who received different scores (e.g., correct versus incorrect) to understand the relationship between student performance on an item and the problem-solving processes and strategies they exhibited.

We hypothesize that the students whose responses received a higher score would apply problem-solving strategies that are more efficient and would engage in more metacognitive behaviors such as reviewing an answer. In contrast, we hypothesize that the lower-scoring students would apply strategies that are less efficient and spend less time engaging in metacognitive behaviors such as planning and monitoring. For example, we predicted that students who received a higher score on an item would solve the problem with fewer D&D response actions, be less likely to revise their answers, and spend less time responding to the item (i.e., higher efficiency) than others who performed less well on the item. Additionally, they would be more likely to apply D&D response strategies that are more efficient on the items (e.g., be more likely to use a target-focused strategy than a source-focused strategy on the G4 item since it is more efficient on this item) than their counterparts who received a lower score. We also predicted that students who solved a problem correctly would allocate more time to the last pause (time elapsed between the last response action and exiting the item), which might be related to metacognitively reviewing the formed response and reflecting on the solution process.

NAEP 2017 mathematics items

This study analyzed process data collected from student interactions with two released items from the 2017 NAEP mathematics assessments. One of these items was administered to fourth graders and the other item was administered to eighth graders. Both items were D&D items.

Grade-Four item

The Grade-Four (G4) item used in the current research (see Fig. 2) evaluates students' knowledge and skills on the mathematical content area Number Properties and Operations, specifically their ability related to the NAEP Mathematics Framework objectives:

(1) Connecting models, number words or numbers using various models and representations for whole numbers, fractions, and decimals, and (2) Representing numbers using models such as base 10 representation, number lines, and two-dimensional models. On this item, test-takers were instructed to drag the decimal numbers (i.e., numeric representations) from the sources and drop them into the targets to denote the value shown in the two-dimensional models. In order to solve this item, they needed to fill each target (t1–t3) with a source (s1–s5). A minimum of three D&D actions were required for a complete and correct response. Students could revise their responses by clicking on the Clear Answer button to remove all objects that had been selected and dropped or by moving a source from a target back to its origin or to another target. Detailed logs of the D&D actions (e.g., add s2 to t1, remove s2 from t1), as well as the corresponding timestamps of these (and other) actions were recorded in process data and used for analysis. An on-screen calculator was not available to students for use on this item.

Grade-Eight item

In the Grade-Eight (G8) item used in this study (see Fig. 3), test-takers were asked to arrange a given set of digits to produce two factors that multiply to a given product. It assesses eighth-grade students' problem-solving skills on the content area Number Properties and Operations, specifically their ability related to the NAEP Mathematics Framework objective: Performing computations with rational numbers. Understanding the inverse relationship between multiplication and division, as well as the multiplication algorithm and its use in problem solving, is expected to help students derive the solution to this problem. Similar to the G4 item, test-takers could form and revise their responses by dragging the numbers in the sources and dropping them into the targets, moving a source from a target back to its original location or to another target, and clicking on the Clear Answer button to remove all objects that had been selected and dropped. Each of the four sources (s1–s4) needs to be dropped to fill the top three-digit factor (t1–t3) and the bottom single-digit factor (t4) to complete the calculation and obtain the given product. A minimum of four D&D actions were required for a complete response on this item. An on-screen scientific calculator was available for use to test-takers upon the click of a Calculator icon provided on the system tool bar. Analysis of calculator use on this item is beyond the scope of the current study.

Methods

Participants

Data for this study were collected from a nationally representative sample of fourth- and eighth-grade students in the U.S. who took the NAEP mathematics assessment administered in 2017 and completed the items listed above (not all NAEP participants took the same items). In this administration, students were asked to complete two mathematics test blocks (they were given 30 min to complete each block) on a handheld tablet. Participants who did not reach or who omitted the items used in the current study and those whose process data on the items were not properly captured were excluded from analysis. A smaller percentage of the fourth graders did not reach the G4 item (0.3%, n = 98) than the percentage of eighth graders who did not reach the G8 item (1.3%, n = 418), given the relative position of the items (G4 item is the first item in a 14-item test block;

G8 item is the fourth item in a 19-item test block). In addition, students who reached the item but did not attempt to fill all the targets in the item (98 fourth graders and 68 eighth graders) were excluded under the premise that this indicated a lack of engagement or a lack of understanding of the directions. In total, 28,385 fourth-grade students who completed the G4 item were included for analysis. Fifty-one percent of the participants self-identified as males (n = 14,523) and 49% of them self-identified as females (n = 13,862). A total of 29,504 eighth-grade students completed the G8 item and were included for analysis. Males comprised 52% of the students (n = 15,224) and females comprised 48% (n = 14,280).

Measures

A list of measures was developed and generated from the process data to infer students' problem-solving processes and strategies. These measures were later combined with the outcome scores on the items to understand the various cognitive and metacognitive processes students who received different scores engaged in and how they responded to mathematics items using different problem-solving strategies.

Score

For the G4 item, test-takers received a full score of 2 if their response was correct (i.e., all three decimals were correctly placed). A partial score of 1 was assigned if two decimals were correctly connected to the two-dimensional models in the response. All other responses where fewer than two decimals were correctly placed were labeled as incorrect and did not receive any credit (score = 0). Among the fourth graders in this study, 57.2% (n = 16,226) of the students received a full score of 2 on this item, 20.9% (n = 5927) of the students received a partial score, while 22.0% (n = 6232) of them did not receive credit.

For the G8 item, test-takers received a full score of 1 if they filled all four digits correctly (i.e., placed 612 and 7 as the two factors). All other responses were incorrect and were assigned a score of 0. On average, 79.3% of the eighth-grade students (n = 23,383) correctly solved this item by identifying both the three-digit factor and the single-digit factor correctly for the given product, showing evidence of their ability to perform computations of whole numbers.

Response action sequences

Sequences of actions executed by students when they responded to each item were extracted from the raw log data and the common response sequences with high frequencies were examined. Several measures were developed from the action sequences in order to understand students' response processes and strategies.

Response sequence length

As mentioned above, students could form and revise their answers by dragging a source into a target, moving the source from a target back to its origin or to another target, or clicking on the Clear Answer button to clear all objects that had been selected and dropped. The number of these aforementioned response-related actions executed by students to form their responses on each item was calculated as an indication of problem-solving efficiency and whether they changed their answers or not (Arslan et al. 2020). If the length of a response sequence was longer than the minimum number of actions required for a complete response on each item (i.e., three actions on the G4 item and four actions on the G8 item), the students had changed their answer at least once by either (re)moving an object they previously dropped or clearing their answer. On average, students executed 4.00 ($SD = 2.16$, $Median = 3$) actions to respond to the G4 item, and an average of 10.29 actions ($SD = 10.74$, $Median = 6$) on the G8 item.

Answer change behaviors

For students who had ever revised their responses (i.e., those whose response sequence was longer than the minimum number of actions required for a complete response on the item), we further examined the patterns of answer change and distinguished the students who changed a correct response to incorrect (correct–incorrect) from those who changed from an incorrect response to a correct one (incorrect–correct). Specifically, students whose initial D&D actions on the targets were not correct (i.e., not selecting all three (on G4 item) or four (on G8 item) correct source objects into the targets with their initial drops to each target), but whose final response was correct (i.e., score = 2 on the G4 item and score = 1 on the G8 item) were labeled as showing an incorrect-correct answer change pattern. In contrast, a correct-to-incorrect pattern was defined as sequences where the correct selections had appeared in the response, but the final response submitted did not receive a full score. Correct-to-correct and incorrect-to-incorrect patterns are not the focus of this analysis.

Classification of initial response strategies

In addition to sequence length and answer change patterns, we classified test-takers' response sequences based on whether the D&D actions were focused on the sources or the targets in order to better understand students' problem-solving strategies. Prior to the classification, the response sequences were cleaned by only keeping the first completed D&D action on each target. Incomplete actions (i.e., started to drag a source but immediately dropped it to its original location) and later revision actions (e.g., moving a dropped source from a target to its origin, or dragging a source to a target that had been previously filled, clicking on the Clear Answer button) were removed from response sequences for strategy classification. In other words, the strategy classification was based on the initial completed D&D action on each target, with an intention to infer the patterns of test-takers' initial attempts to solve a problem. The cleaned sequences of response-related events on each item were then classified into four categories (source-focused, target-focused, mixed, and indistinguishable) based on the following definitions. This classification was followed and applied in our later work (Arslan et al. 2020).

Source-focused strategy In this strategy, test-takers focus on the source objects in an item and drag the sources sequentially (either in ascending or descending order) into the corresponding targets. An example sequence showing a source-focused strategy on the G4 item is: dragging *s1* to t2; *s2* to t1; and *s3* to t3, where the sources are dragged in ascending order. Note that we only consider the sequences of the first D&D events on each target for strategy identification.

Target-focused strategy A target-focused strategy is defined as filling targets sequentially (in either ascending or descending order). An example sequence showing a target-focused strategy on the G4 item is: dragging s2 to *t1*; s1 to *t2*; and s3 to *t3*. In this example, the test-taker focuses on t1 and drops the matching source into this target before moving to t2, where they repeat the same procedure and move on to t3. Note that the four targets in the G8 item were not placed horizontally as in the G4 item. On this item, the multiplication problem was written vertically and the target for the single-digit factor was placed vertically below the three-digit factor (specifically t3). Therefore, response sequences where test-takers filled the bottom factor (t4) first and then filled the targets in the upper factor from left to right (e.g., s4 to *t4*; s3 to *t1*; s1 to *t2*; and s2 to *t3*) and sequences where the upper factor was filled from right to left before filling the bottom factor (e.g., s2 to *t3;* s1 to *t2*; s3 to *t1*; s4 to *t4*) were also classified as a target-focused strategy.

Indistinguishable strategy We labeled sequences where sources are dragged sequentially (in either ascending or descending order) into the targets in sequential order (e.g., s1 to t1; s2 to t2; s3 to t3, or s5 to t3; s3 to t2; s1 to t1 on the G4 item) as indistinguishable because we could not distinguish whether a test-taker is consciously displaying a source-focused or a target-focused strategy, or these selections were made due to disengagement or random guessing. Note that these action sequences could not lead to a correct response on either item unless revisions were made after the initial attempts. Therefore, it is our expectation that these sequences were significantly less frequent among the students who provided a correct response compared to others.

Mixed strategy All other response sequences that do not fall into the source-focused, target-focused, or indistinguishable categories and therefore do not follow a systematic pattern are classified as a mixed strategy.

As shown in Fig. 4, 56.8% (n = 16,116) of the fourth-grade test-takers adopted a target-focused strategy and filled the targets sequentially; 17.8% of the students (n = 5053) focused on the sources and dragged and dropped the sources sequentially; 10.3% (n = 2926) of the response sequences were not distinguishable between a source-focused and a target-focused strategy; and 15.1%, (n = 4290) of the sequences did not show a clear pattern. On the G8 item, 59.8% (n = 17,646) of the eighth-grade test-takers used a target-focused strategy; 6.1% (n = 1792) dragged the sources sequentially (i.e., a source-focused strategy); and the remaining sequences showed unsystematic patterns (20.2%, n = 5955) or were not distinguishable between source-focused and target-focused strategies (13.9%, n = 4111).

Time

The time-based measures computed in this study included the total response time, first pause, total D&D execution time, average D&D execution time, and last pause. Specifically, response time is the total amount of time spent on each item (i.e., item completion time). Similar to our later work (Arslan et al. 2020), first pause is defined as the time between the item first appearing on the screen and the test-taker's first D&D action. It is representative of the time needed to encode information in the problem stem, mentally represent the item, conduct necessary mental computations for problem solving, and make decisions on the first D&D event. Similarly, last pause is the time elapsed

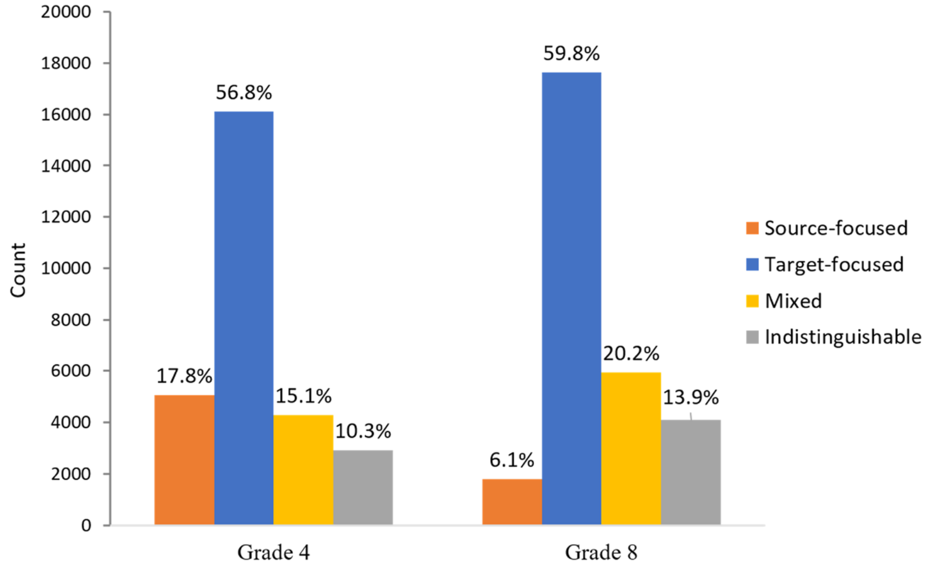


Fig. 4 Number of test-takers who applied a source-focused, target-focused, mixed, or indistinguishable strategy in responding to the fourth- and eighth-grade items

between the last D&D action and the item last appearing on the screen, possibly indicating that test-takers were reviewing a solution they had formed. Total D&D execution time is the time elapsed between the first and the last D&D actions. This reflects the time needed to perform necessary mental calculations, make decisions on the following response actions, and execute these actions. Considering that total D&D execution time is associated with the number of D&D actions executed, we also computed average execution time per response action by dividing the total D&D execution time by (response sequence length – 1). Average D&D execution time measures the average transition time between two consecutive response actions and probably reveals the efficiency of problem-solving. Each of these measures (first pause, total D&D execution time, average D&D execution time, and last pause) was computed as both the absolute values in minutes and as the proportion out of the total time spent on the item, resulting in a total of nine time-based measures (see the full list in Tables 3 and 5).

Note that students had the flexibility to move freely among items in each timed block and exit an item to work on other items and revisit it at any time. Time spent on other items was not included in any time-based measures used in this study. A 90% winsorization was applied to all time-related measures on each item to exclude extreme cases that possibly represent off-task behaviors or issues in timestamp logging. In this process, the top 5% extreme values of each measure were replaced by the value at the 95th percentile, and the bottom 5% extreme values were replaced by the value at the 5th percentile.

Data analysis

In this paper, statistical tests were conducted to compare the aforementioned process-related measures representative of test-takers' cognitive and metacognitive processes and strategies across different score groups. Specifically, Chi-square tests were conducted on the categorical variables (e.g., response strategy groups) to investigate the potential differences in the problem-solving processes between students of different scores. Odds ratio (OR) was obtained and reported as a measure of effect size for Chi-square tests. As the continuous measures such as response sequence length and response time were not normally distributed, two-tailed Mann–Whitney U tests, a non-parametric alternative to the t-test, were conducted to compare these measures between the students who answered the G8 item correctly and those who answered incorrectly. The null hypothesis of a Mann–Whitney U test is that the probability of a randomly selected value from the first population being greater than a randomly selected value from the second population is 50%. As a measure of effect size, point biserial correlation r was obtained for Mann–Whitney tests (Fritz et al. 2012). For the G4 item, omnibus Kruskal–Wallis tests and pairwise Mann–Whitney U tests were conducted to compare the continuous measures across different score groups. Given the substantial number of statistical tests, we controlled for the proportion of false positives by applying Benjamini and Hochberg's (1995) False Discovery Rate post-hoc method.

Results

Grade-Four Item

Response action sequences

Table 2 presents the most frequent response sequences executed by students who scored differently on the item, and the frequency and proportion of these sequences within each score group. Among the students who correctly connected all three decimal numbers with the two-dimensional models, the most common sequence was exhibited by nearly half of the students (49.6%) and involved dragging the correct sources into t1, t2, and t3 in order (a target-focused strategy). Other frequent response sequences for students who received a full credit but used a "less ideal" path with more than three actions involved either clearing a previously entered correct answer and making the same three drag-and-drops again, or changing a previously dropped incorrect answer (e.g., s1-to-t1 or s5-to-t3) and immediately replacing it with a correct one. On the other hand, common response sequences leading to incorrect or partially correct solutions mostly involved errors of dragging the sources related to the decimal numbers 2.0 and 2.5 incorrectly into the targets (e.g., dropping s4, number 2.0, into t1 or t2, or dropping s5, number 2.5, into t3).

Response sequence length

On average, students who received full credit (score of 2) exhibited significantly shorter D&D response sequences ($M = 3.88$, $SD = 1.92$, $Median = 3$) than their counterparts who received partial credit (score of 1, $M = 4.02$, $SD = 2.00$, $Median = 3$), $U = 51{,}004{,}947.5$, $p < 0.001$, $r = 0.06$. Similarly, the students who received partial credit executed significantly fewer D&D actions than students who received no credit (score of 0, $M = 4.28$, $SD = 2.78$, $Median = 3$), $U = 19{,}621{,}889.5$, $p < 0.001$, $r = 0.06$. It is important to note that the effect sizes for these comparisons were very small and the significant findings could be simply an effect of the large sample size.

A total of 19,474 (68.6%) fourth-grade test-takers executed exactly three D&D actions to form their responses, which is the lowest number of actions required for

Table 2 Top ten most frequent response sequences among students who received different scores on the G4 item and their frequency and proportion within each score group

Score	Response Action Sequence	Freq	Pct
0 (n = 6232)	Add_s4_t1; Add_s1_t2; Add_s5_t3	644	10.3%
	Add_s2_t1; Add_s4_t2; Add_s5_t3	274	4.4%
	Add_s1_t2; Add_s4_t1; Add_s5_t3	215	3.4%
	Add_s1_t2; Add_s5_t3; Add_s4_t1	214	3.4%
	Add_s1_t1; Add_s4_t2; Add_s3_t3	198	3.2%
	Add_s1_t1; Add_s4_t2; Add_s5_t3	181	2.9%
	Add_s4_t1; Add_s5_t3; Add_s1_t2	129	2.1%
	Add_s5_t3; Add_s1_t2; Add_s4_t1	113	1.8%
	Add_s5_t3; Add_s4_t1; Add_s1_t2	100	1.6%
	Add_s1_t1; Move_s1_t2; Add_s4_t1; Add_s5_t3	75	1.2%
1 (n = 5927)	Add_s2_t1; Add_s1_t2; Add_s5_t3	1141	19.3%
	Add_s2_t1; Add_s4_t2; Add_s3_t3	1022	17.2%
	Add_s1_t2; Add_s2_t1; Add_s5_t3	286	4.8%
	Add_s4_t1; Add_s1_t2; Add_s3_t3	210	3.5%
	Add_s4_t2; Add_s2_t1; Add_s3_t3	207	3.5%
	Add_s1_t2; Add_s5_t3; Add_s2_t1	163	2.8%
	Add_s4_t2; Add_s3_t3; Add_s2_t1	146	2.5%
	Add_s2_t1; Add_s3_t3; Add_s4_t2	97	1.6%
	Add_s5_t3; Add_s1_t2; Add_s2_t1	96	1.6%
	Add_s2_t1; Add_s5_t3; Add_s1_t2	78	1.3%
2 (n = 16,226)	Add_s2_t1; Add_s1_t2; Add_s3_t3	8043	49.6%
	Add_s1_t2; Add_s2_t1; Add_s3_t3	2098	12.9%
	Add_s1_t2; Add_s3_t3; Add_s2_t1	792	4.9%
	Add_s2_t1; Add_s3_t3; Add_s1_t2	565	3.5%
	Add_s2_t1; Add_s1_t2; Add_s3_t3; Clear Answer; Add_s2_t1; Add_s1_t2; Add_s3_t3	269	1.7%
	Add_s1_t1; Move_s1_t2; Add_s2_t1; Add_s3_t3	253	1.6%
	Add_s3_t3; Add_s1_t2; Add_s2_t1	216	1.3%
	Add_s2_t1; Add_s1_t2; Add_s5_t3; Rem_s5_t3; Add_s3_t3	149	0.9%
	Add_s2_t1; Add_s1_t2; Add_s3_t3; Clear Answer; Add_s1_t2; Add_s2_t1; Add_s3_t3	143	0.9%
	Add_s3_t3; Add_s2_t1; Add_s1_t2	123	0.8%

Add_s1_t2 represents dragging source 1 into target 2; Rem_s1_t1 represents removing source 1 from target 1 back to its original location; Move_s1_t2 represents moving source 1 from a previous target to target 2.

a complete solution and thus the most efficient. The remaining 31.4% (n = 8911) of the students showed response sequences longer than three actions, indicating that they changed their answers during the problem-solving process. Students who received full credit were more likely to execute exactly three actions for responses (73.0%) compared to those who received partial credit on the item (66.2%), $\chi^2(1, N = 22,153) = 97.31$, $p < 0.001$, $OR = 1.38$, and those who received no credit (59.6%), $\chi^2(1, N = 22,458) = 374.73$, $p < 0.001$, $OR = 1.83$. The difference in the proportion of three-action response sequences between the students who received a score of 0 and those who received a score of 1 was also statistically significant, $\chi^2(1, N = 12,159) = 55.13$, $p < 0.001$, $OR = 1.32$.

Answer change

Among the fourth graders who had changed their responses (i.e., whose response sequences were longer than three), 32.0% (n = 2853) of them did not make all three correct connections of the representations initially, but eventually revised their response into a correct one. Of these incorrect-to-correct students, 485 revised their answers from incorrect (score = 0) to correct (score = 2), and 2368 revised their responses from partially correct (score = 1) to correct. In contrast, fewer students (2.0%, n = 179) changed answers from correct to incorrect or partially correct. Sixty of them had all three decimals correctly placed in their response sequences, but received a final score of 0 based on their response. One hundred and nineteen students had made the correct connections but eventually revised their response into partially correct. Students who changed response from incorrect or partially correct to correct comprised 17.6% of those who received a full score. Students who changed responses from correct to incorrect or partially correct comprised 1.5% of the students who did not receive full credit on the item.

Response strategies

Among the students who received full credit, 64.5% (n = 10,457) adopted a target-focused strategy to give a response (see Fig. 5); 18.2% (n = 2956) focused on the sources and dragged them sequentially; 4.3% (n = 695) of the response sequences were indistinguishable between a source- and a target-focused strategy; while 13.1% (n = 2118) of the sequences were in the *Mixed* category. In contrast, only 35.2% (n = 2192) of the students who did not receive any credit on the item used a target-focused approach; 17.2% (n = 1072) of them focused on the sources; 27.3% of the sequences were indistinguishable and 20.3% (n = 1266) of the sequences were mixed. Among the students who received a partial credit, 58.5% (n = 3467) used a target-focused strategy, 17.3% (n = 1025) applied a source-focused strategy, 8.9%

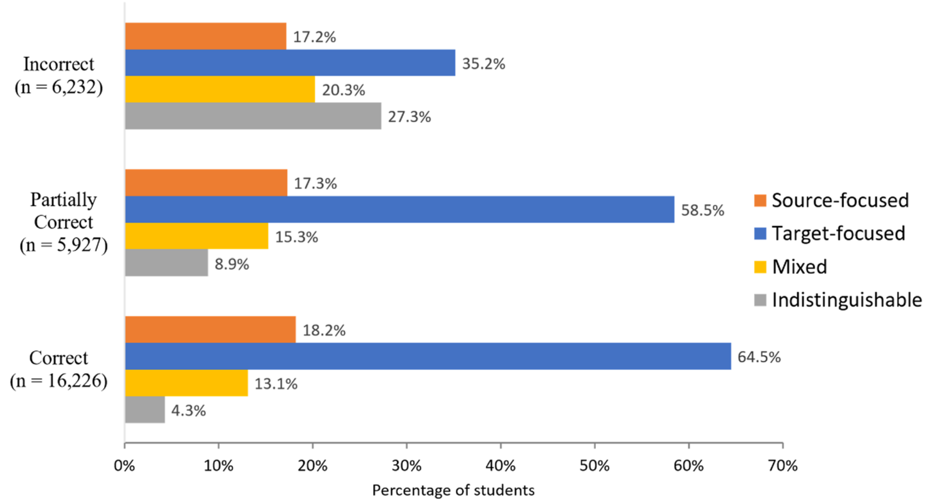


Fig. 5 Percentage of fourth-grade students in each score group who applied a source-focused, target-focused, indistinguishable, or mixed strategy in response to the item

(n = 529) executed sequences that were indistinguishable between strategies, and 15.3% (n = 906) did not show a systematic pattern in their response sequences (see Fig. 5). Chi-square tests suggested that the students whose answer received partial credit and those whose answer received full credit were significantly more likely than the students whose answer received no credit to adopt a target-focused strategy over a source-focused strategy ($\chi^2(1, N = 7,756) = 95.80$, $p < 0.001$, $OR = 1.65$; $\chi^2(1, N = 16,677) = 166.71$, $p < 0.001$, $OR = 1.73$). The odds of using a mixed strategy over a target-focused strategy or over a source-focused strategy by the students whose answer received no credit was significantly higher than the odds for the partial-scoring students ($\chi^2(1, N = 7,831) = 242.55$, $p < 0.001$, $OR = 2.21$; $\chi^2(1, N = 4,269) = 21.83$, $p < 0.001$, $OR = 1.34$) and the full-scoring students ($\chi^2(1, N = 16,033) = 635.28$, $p < 0.001$, $OR = 2.85$; $\chi^2(1, N = 7,412) = 98.79$, $p < 0.001$, $OR = 1.65$). In addition, results indicated that the indistinguishable sequences were significantly less frequent among the students who provided a correct response compared to the others. This is expected because these indistinguishable sequences could not lead to a correct response unless revisions were made after the initial attempts.

Time

Students whose answer received full credit spent significantly less time (in minutes) on this item (see Table 3; $M = 1.45$, $SD = 0.66$, $Median = 1.29$) than the students whose response received no credit ($M = 1.52$, $SD = 0.72$, $Median = 1.34$, $U = 52,609,232$, $p < 0.001$, $r = –0.03$) and those whose response received partial credit ($M = 1.57$, $SD = 0.70$, $Median = 1.39$, $U = 52,785,444.5$, $p < 0.001$, $r = – 0.08$). Further breakdown of the response time indicated that the first pause constituted a significantly smaller proportion of their total response time for the students who received full credit ($M = 54.0\%$, $SD = 15.6\%$, $Median = 53.9\%$) than those who received no credit ($M = 56.7\%$, $SD = 17.0\%$, $Median = 57.3\%$, $U = 55,526,457$, $p < 0.001$, $r = – 0.08$) and those who received partial credit ($M = 55.2\%$, $SD = 15.9\%$, $Median = 55.4\%$, $U = 50,358,868.5$, $p < 0.001$, $r = – 0.04$). On the contrary, students who received full credit distributed a significantly larger proportion of their time on the last pause ($M = 10.7\%$, $SD = 8.6\%$, $Median = 8.0\%$) than those whose answer received partial credit ($M = 9.7\%$, $SD = 8.0\%$, $Median = 7.2\%$, $U = 44,852,316$, $p = 0.009$, $r = 0.05$) and those whose answer received no credit ($M = 9.6\%$, $SD = 8.1\%$, $Median = 7.0\%$, $U = 46,723,640.5$, $p < 0.001$, $r = 0.06$). Similar trends were found for the time measures in their absolute values. On the other hand, students who did not receive any credit showed significantly shorter transition time when they transitioned between D&D response actions (i.e., average execution time) ($M = 11.6\%$, $SD = 6.0\%$, $Median = 10.4\%$) than their counterparts who received full credit ($M = 13.6\%$, $SD = 5.9\%$, $Median = 13.0\%$, $U = 40,371,216.5$, $p < 0.001$, $r = – 0.16$) and those who received partial credit ($M = 13.0\%$, SD = 6.0%, $Median = 12.1\%$, $U = 15,857,103.5$, $p < 0.001$, $r = – 0.12$). Similar results were obtained for the total D&D execution time. Note that the effect sizes for the comparison of the time measures were all relatively small and the significant results could be simply caused by the large sample size.

Table 3 Descriptive statistics of the process-related measures on the G4 item by score group

Type	Measure	Incorrect			Partially Correct			Correct			*P*
		Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	
Sequence	Response sequence length	4.28	3.00	2.78	4.02	3.00	2.00	3.88	3.00	1.92	< 0.001
Time	Total response time	1.52	1.34	0.72	1.57	1.39	0.70	1.45	1.29	0.66	< 0.001
	First pause	0.86	0.71	0.52	0.87	0.72	0.51	0.80	0.66	0.48	< 0.001
	Pct. first pause	56.7%	57.3%	17.0%	55.2%	55.4%	15.9%	54.0%	53.9%	15.6%	< 0.001
	Total D&D execution time	0.47	0.38	0.31	0.51	0.43	0.30	0.48	0.40	0.28	< 0.001
	Pct. total D&D execution time	32.7%	30.5%	15.8%	34.2%	32.5%	14.9%	34.3%	32.8%	14.3%	< 0.001
	Avg. D&D execution time	0.16	0.13	0.09	0.18	0.16	0.09	0.18	0.16	0.09	< 0.001
	Pct. avg. D&D execution time	11.6%	10.4%	6.0%	13.0%	12.1%	6.0%	13.6%	13.0%	5.9%	< 0.001
	Last pause	0.13	0.09	0.11	0.13	0.10	0.11	0.14	0.10	0.11	< 0.001
	Pct. last pause	9.6%	7.0%	8.1%	9.7%	7.2%	8.0%	10.7%	8.0%	8.6%	< 0.001

P-values of omnibus Kruskal–Wallis tests are reported. Pairwise Mann–Whitney U tests were also conducted and discussed in the text

Grade-Eight item

Response action sequences

Table 4 presents the most frequent response sequences among the eighth graders who answered the G8 item correctly and incorrectly. Based on the results, the most common sequence in attempting to solve this problem among the correct respondents (adopted by 35.1% of these students) was filling the top three-digit factor correctly from left to right, and dragging the number 7 (s4) and dropping it into the target for the bottom single-digit factor (t4). A second most frequent pattern (6.5%) leading to a correct answer involves filling the bottom single-digit factor before moving on to fill the hundreds, tens, and ones places of the three-digit factor in order. Both sequences were classified as a target-focused strategy. On the other hand, the most common response sequence leading to an incorrect answer was dragging s1 and dropping it to t1, then s2 to t2, s3 to t3, and s4 to t4.

Response sequence length

Similar to the G4 item, students who successfully solved the G8 problem executed significantly fewer actions ($M = 9.70$, $SD = 9.86$, $Median = 5$) on average to form their responses than students who answered the item incorrectly ($M = 12.52$, $SD = 13.37$, $Median = 7$), $U = 82{,}157{,}780$, $p < 0.001$, $r = 0.11$. In total, 46.6% (n = 13,738) of the eighth-grade students completed the item with four D&D actions, the minimum number of

Table 4 Top ten most frequent response sequences among students who received different scores on the G8 item and their frequency and proportion within each score group

Score	Response Action Sequence	Freq	Pct
0 (n = 6121)	Add_s1_t1; Add_s2_t2; Add_s3_t3; Add_s4_t4	432	7.1%
	Add_s3_t1; Add_s2_t2; Add_s1_t3; Add_s4_t4	231	3.8%
	Add_s4_t1; Add_s1_t2; Add_s2_t3; Add_s3_t4	231	3.8%
	Add_s4_t1; Add_s2_t2; Add_s1_t3; Add_s3_t4	140	2.3%
	Add_s2_t1; Add_s1_t2; Add_s3_t3; Add_s4_t4	59	1.0%
	Add_s4_t4; Add_s3_t1; Add_s2_t2; Add_s1_t3	49	0.8%
	Add_s4_t1; Add_s3_t2; Add_s1_t3; Add_s2_t4	44	0.7%
	Add_s1_t1; Add_s3_t2; Add_s2_t3; Add_s4_t4	41	0.7%
	Add_s2_t1; Add_s3_t2; Add_s1_t3; Add_s4_t4	39	0.6%
	Add_s1_t1; Add_s3_t2; Add_s4_t3; Add_s2_t4	32	0.5%
1 (n = 23,383)	Add_s3_t1; Add_s1_t2; Add_s2_t3; Add_s4_t4	8219	35.1%
	Add_s4_t4; Add_s3_t1; Add_s1_t2; Add_s2_t3	1509	6.5%
	Add_s3_t1; Add_s2_t3; Add_s1_t2; Add_s4_t4	476	2.0%
	Add_s4_t4; Add_s2_t3; Add_s1_t2; Add_s3_t1	326	1.4%
	Add_s2_t3; Add_s4_t4; Add_s1_t2; Add_s3_t1	228	1.0%
	Add_s3_t1; Add_s1_t2; Add_s4_t4; Add_s2_t3	168	0.7%
	Add_s4_t4; Add_s3_t1; Add_s2_t3; Add_s1_t2	124	0.5%
	Add_s1_t2; Add_s2_t3; Add_s3_t1; Add_s4_t4	105	0.4%
	Add_s1_t1; Rem_s1_t1; Add_s3_t1; Add_s1_t2; Add_s2_t3; Add_s4_t4	89	0.4%
	Add_s3_t1; Add_s1_t2; Add_s2_t3; Add_s4_t4; Clear Answer; Add_s3_t1; Add_s1_t2; Add_s2_t3; Add_s4_t4	83	0.4%

Add_s1_t2 represents dragging source 1 into target 2; Rem_s1_t1 represents removing source 1 from target 1 back to its original location; Move_s1_t2 represents moving source 1 from a previous target to target 2 actions required for a complete response on the item. Specifically, 49.4% (n = 11,551) of the eighth graders who answered the item correctly formed their solution with exactly four D&D steps without revising their response, while 35.7% (n = 2187) of those who did not provide a correct answer used exactly four actions to form a solution. A Chi-square test indicated that the difference between the correct and incorrect score groups was statistically significant, $\chi^2(1, N = 29,504) = 363.76$, $p < 0.001$, $OR = 1.76$.

Answer change

Among the eighth-grade students who changed their answers, only 0.6% (n = 95) of them had identified all the four digits correctly based on their sequences but revised their answers and eventually submitted an incorrect response to the item. These students comprised 1.6% of the students in the incorrect score group. On the other hand, 71.3% (n = 11,235) of the students who revised their answers did not make all four correct drag-and-drops right away within the first four response actions but submitted a correct final answer. These students comprised 48.0% of the correct respondents.

Response strategies

Among the students whose response was correct (see Fig. 6), more students adopted a target-focused strategy (64.4%, n = 15,048) or a mixed approach (19.0%, n = 4449) than a source-focused strategy (5.1%, n = 1197). Similarly, target-focused strategy (42.4%, n = 2598) and mixed strategy (24.6%, n = 1506) were more common than source-focused strategy (9.7%, n = 595) among the students whose response was incorrect (see Fig. 6). Chi-square tests indicated that the odds of using a target-focused strategy and a mixed strategy as opposed to a source-focused strategy were significantly higher for the students who answered the item correctly than those who answered incorrectly ($\chi^2(1, N = 19,438) = 403.35$, $p < 0.001$, $OR = 2.99$; $\chi^2(1, N = 7,747) = 43.24$, $p < 0.001$, $OR = 1.47$). Students who submitted a correct answer were also significantly more likely to focus on the targets over adopting a mixed approach than those who submitted an incorrect answer ($\chi^2(1, N = 23,601) = 345.34$, $p < 0.001$, $OR = 1.96$).

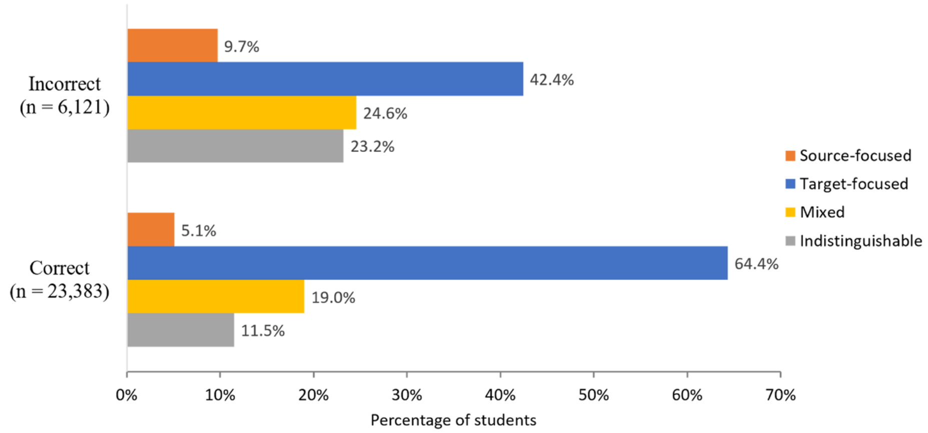


Fig. 6 Percentage of eighth-grade students with a correct and incorrect response who applied a source-focused, target-focused, indistinguishable, or mixed strategy in responding to the item

Time

As shown in Table 5, students whose answer was correct spent significantly less time (in minutes) on this item ($M=2.20$, $SD=1.29$, $Median=1.84$) than students who answered the item incorrectly ($M=2.66$, $SD=1.70$, $Median=2.31$), $U=79{,}458{,}658$, $p<0.001$, $r=-0.08$. Further breakdown of the response time indicated that the students who solved the problem correctly distributed a significantly larger proportion of their response time on the first pause than those who answered incorrectly ($Ms=56.2\%$ and 47.6%, $Medians=65.3\%$ and 44.8%, $U=60{,}635{,}405.5$, $p<0.001$, $r=0.11$). It also took the students who solved the problem correctly significantly longer in their last pause (i.e., the transition between the last response action and exiting the item) ($Ms=8.6\%$ and 8.4%, $Medians=5.4\%$ and 4.6%), $U=66{,}284{,}674$, $p<0.001$, $r=0.05$. On the other hand, students who solved the problem correctly spent significantly less time in executing and completing the response actions (both in terms of the total and average execution time) than their counterparts who failed to correctly solve the item ($U=83{,}603{,}381$, $p<0.001$, $r=-0.12$; $U=81{,}234{,}841$, $p<0.001$, $r=-0.11$). Note that the effect sizes for the comparison of the time measures were relatively small.

Discussion and conclusion

In this exploratory study, we used process data to attempt to understand the cognitive and metacognitive processes that fourth and eighth graders in the United States engaged in on two technology-enhanced items in NAEP 2017 mathematics assessment. Specifically, measures were developed and generated from process data to characterize students' problem-solving strategies and their allocation of time during the response processes. Results from this research revealed that test-takers who achieved a higher level of accuracy on an item applied problem-solving strategies that were more efficient when responding to the D&D items. They also spent more time engaging in metacognitive behaviors such as reviewing a previously submitted solution. On the contrary, the lower-scoring students tended to use strategies that were less efficient when attempting to solve the item and spent less time engaging in metacognitive monitoring behaviors. These results not only showed validity evidence of the item scores, but also added to the

Table 5 Descriptive statistics of the process-related measures on the G8 item by score group

Type	Measure	Incorrect			Correct			*r*	*P*
		Mean	Median	SD	Mean	Median	SD		
Sequence	Response sequence length	12.52	7.00	13.37	9.70	5.00	9.86	0.11	< 0.001
Time	Total response time	2.66	2.31	1.70	2.20	1.84	1.29	–0.08	< 0.001
	First pause	1.14	0.54	1.17	1.14	0.80	1.01	0.06	< 0.001
	Pct. first pause	47.6%	44.8%	32.1%	56.2%	65.3%	34.7%	0.11	< 0.001
	Total D&D execution time	1.16	0.53	1.28	0.83	0.23	1.06	–0.15	< 0.001
	Pct. total D&D execution time	42.2%	38.5%	31.1%	34.4%	17.1%	32.7%	–0.12	< 0.001
	Avg. D&D execution time	0.09	0.07	0.08	0.08	0.05	0.07	–0.13	< 0.001
	Pct. avg. D&D execution time	4.8%	3.8%	3.6%	3.9%	2.7%	3.3%	–0.11	< 0.001
	Last pause	0.15	0.07	0.17	0.15	0.10	0.14	0.06	< 0.001
	Pct. last pause	8.4%	4.6%	9.2%	8.6%	5.4%	8.4%	0.05	< 0.001

P-values of Mann–Whitney U tests and the effect sizes of the comparisons (*r*) are reported.

outcome scores by providing a deeper understanding of how the solutions were reached. Findings also provide insights into fourth and eighth graders' common misconceptions on mathematics topics and where test-takers may have struggled in the problem-solving process.

Response action sequences

In general, results indicated that students who answered the mathematics items correctly tended to also solve them in a more efficient way. For test-takers at each grade, those who received a higher score on an item solved the problem with significantly fewer D&D response steps and used significantly shorter response time than their counterparts who performed less well on the item. A significantly larger proportion of the test-takers who correctly answered each item formed their solutions without clearing or revising them than their counterparts who performed less well on the item, suggesting their higher proficiency and confidence in the knowledge component assessed. By contrast, students who provided an incorrect answer, probably due to their lack of knowledge, were either less confident about their solution and changed their answers back and forth, or took more steps to finalize a solution through strategies such as a trial-and-error approach. This is also consistent with the findings that students who answered the item correctly spent significantly less time to give a response, solving the problem more efficiently than the students who answered incorrectly.

Further examination of the answer change behaviors indicated that the fourth and eighth graders who changed their answers were more likely to revise their responses from incorrect to correct and therefore made score gains on the item. This suggests that findings from previous literature that test-takers benefit from changing answers on MC items in paper-based assessments (Al-Hamly and Coombe 2005; Bauer et al. 2007) and on MC items in computer-based assessments (Liu et al. 2015; Mcconnell et al. 2012) also apply to the more interactive D&D items in technology-enhanced assessments.

The frequent response sequences executed by test-takers not only shed light on the common misconceptions shown and strategies used by students who had difficulty in providing a correct solution (e.g., why they answered an item incorrectly), but also help distinguish students who submitted the same correct answer but adopted different problem-solving strategies. Below we discuss the problem-solving strategies that were inferred from process data.

On the G4 item, most of the test-takers used a target-focused strategy and filled targets in a sequential manner. Students who correctly connected all three decimals to corresponding two-dimensional models were more likely than the other students who failed to translate the representations correctly to adopt a target-focused strategy than a source-focused strategy and a mixed approach. These students probably focused on the two-dimensional models and converted each of them into a symbolic decimal, and immediately dragged the corresponding decimal into the target before moving on to solve the next target. To them, the models in the targets are to-be-solved/converted mathematical objects while the sources are the symbolic representations to be matched. As we discussed in the Introduction section, applying a target-focused strategy is more efficient than a source-focused strategy in this case, considering the fewer cognitive steps required for students to solve the item and the relatively lower cognitive load involved. A smaller proportion of students, especially among those who received a full score, applied a source-focused strategy on this item. In other words, students who scored higher on the item probably searched their strategy repertoire, evaluated the efficiency of strategies, and decided to adopt a strategy that was the most efficient to give a response.

In addition, students who did not correctly link the symbolic representations with the visual ones were more likely to exhibit unsystematic response sequences that did not show clear patterns or were not distinguishable between strategies, indicating that they might be struggling with problem solving and started with sources or targets that they found the easiest to solve instead of working on them sequentially. It is also possible that these students were simply off-task or randomly guessing (dragging sources into the targets that they had randomly guessed).

Exploration of the sequences executed by students who answered the G4 item incorrectly indicated that the common errors they made involved dragging s5 (i.e., decimal number 2.5) into t3 (two-dimensional model representing 0.25), dragging s4 (decimal number 2.0) into t2 (two-dimensional model representing 0.02) or t1 (two-dimensional model representing 0.20), and dragging s1 (decimal number 0.02) into t1. For instance, the most frequent pattern leading to a score-0 answer was Add_s4_t1; Add_s1_t2; Add_s5_t3. These students filled the targets sequentially (a target-focused strategy), but their response showed a lack of basic conceptual understanding of the representations of decimals and the place value after the decimal point. Their errors could be an indication of a lack of understanding of the whole-part relationship and/or the existence of whole number bias (Ni and Zhou 2005; Resnick et al. 1989; Roche 2010; Westenskow et al. 2014). Whole number bias refers to the tendency to incorrectly apply whole number schemes/rules to interpret fractions or decimal fractions. In this case, students might have focused on counting the quantity of the shaded lines in the model (e.g., 2 in t1) instead of the magnitude of the part-whole relationship.

Analysis of the response sequences executed by the students who gave a correct answer provided insights about how test-takers arrived at the solution and helped distinguish test-takers who adopted an efficient strategy from those who solved problems less efficiently likely because of a lack of knowledge or construct-irrelevant noise. In addition to the patterns where the fourth-grade test-takers formed an answer without any change, the most common sequences among the students who received full credit on the G4 item involved: (1) Clearing a correct answer previously entered and then re-entering the same answer; (2) filling t1 with s1 (decimal 0.02) and correcting the answer immediately by moving the source from t1 to t2 (Add_s1_t1; Move_s1_t2; Add_s2_t1; Add_s3_t3); or (3) removing the number 2.5 from the t3 and dragging the correct source (decimal 0.25) into it (Add_s2_t1; Add_s1_t2; Add_s5_t3; Rem_s5_t3; Add_s3_t3). The first path might be associated with the test-takers' unfamiliarity with the system interface. It is likely that they misused the Clear Answer button to submit their response and redid the same drag-and-drops after realizing the misuse. It might be helpful to provide instructions on how to use the button prior to the test-taking process to minimize the confusion for these students. It is also possible that these students worked the problem for a second time to confirm their previously formed solution. In sequences such as (2) and (3), on the other hand, students dragged an incorrect source into a target, either due to carelessness or the existence of a common misconception, and then revised the answer into a correct one. These sequences should be distinguished from sequences in (1) and sequences where a correct response was formed without any changes even though they led to exactly the same final response and score. For example, if the errors shown in the sequences were not carelessly made, the students might not have fully understood the place value of decimals and how to symbolically and graphically represent decimals. It is important to identify these students with shallow knowledge (either through teacher feedbacks or allocation of partial credit) for future instructions and scaffolding to reinforce a thorough understanding.

Unlike the G4 item where the to-be-filled targets are relatively independent of each other and solving the two-dimensional representation in the first target is relatively independent of solving the second target, the four digits in the G8 problem are interrelated and finding the value of a digit is dependent on the values in the other digits. Students need to consider the four digits as a whole for a response. Compared to the G4 item, a larger proportion of the students' response sequences on the G8 item did not follow a systematic order and belonged to the mixed strategy category. These students did not necessarily focus on the sources or targets. Instead, they might have used a strategy to identify the value for a target that was the easiest to solve, immediately filled it once they reached a decision on this target, and then moved to the next easily solvable target. The mixed strategy might also indicate the use of a trial-and-error approach (Elia et al. 2009) in problem solving. In contrast, fewer students used a source-focused strategy when solving this item.

On the G4 item, most test-takers (68.6%) formed their answers without revising them, while a smaller proportion (46.4%) of the eighth-grade test-takers completed the G8 item without answer revisions. The response sequences were also longer on the G8 item than on the G4 item. One possible explanation is that test-takers needed to change their answers more often on the G8 item, possibly because they were applying strategies such as a trial-and-error approach or a guess-check-revise approach. These strategies are computationally less efficient considering the relatively more computational steps required (e.g., the numerous iterations of guessing involved in the guess-check-revise strategy). Note the difference might also be related to the difference in the minimum number of actions required for a complete response on the G4 and G8 items (3 vs. 4).

The most common sequence among the test-takers who correctly solved the G8 problem was filling the top three-digit factor sequentially, and then dragging s4 and dropping it into the bottom single-digit factor. This is a target-focused strategy. The students might have solved the problem by focusing on the targets sequentially and making decisions on the value in each target and immediately executing the drops after a decision was made on each target. Alternatively, these students might have made decisions on all the four digits through mental computations or with the assistance of tools such as calculators before they started to execute the D&D actions.

A second most frequent pattern among the students who provided a correct solution to the G8 item involves filling the bottom single-digit factor before completing the three-digit factor from left to right. Note that several frequent sequences exhibited by students who answered correctly started with filling the single-digit factor. Further examination revealed that 22.0% of the students who scored correctly started with this digit, which is consistent with the importance of this digit for problem-solving. A mental process applied by these students might be: Considering that the “42” in the product can only be obtained by 6 × 7, and only 2 × 7 can lead the ones place in the product to be 4, the common number 7-should be placed as the single-digit factor. Another reasoning process might be based on recognizing how using the inverse operation of multiplication-division-can assist with solving the problem. In this strategy, the number to be placed in t4 has to be such that a three-digit factor is obtained when 4,284 is divided by this number. There are four options for t4: 1, 2, 6, or 7. Using number sense, the numbers 1 and 2 are inappropriate given that when dividing 4,284 by 1 or 2, the resulting factor is a four-digit number. Thus, the only potential numbers to consider for t4 are 6 and 7. However, the number 6 is ruled out because the only potential units digit of the three-digit factor would be either 4 or 9 in order for the units digit of the product to be 4, none of which is among the sources. Therefore, the only appropriate source for t4 is s4 (number 7). Note that among all students with response sequences that started with s4 in t4 (11.3%), 90.2% answered the item correctly, validating the value of this strategy.

On the other hand, the most frequent response sequence leading to an incorrect answer among the eighth graders was dragging s1 and dropping it into t1, then dragging and dropping s2 to t2, s3 to t3, and s4 to t4. This sequence suggests that students might be engaged in random guessing behaviors because they do not know (Budescu and Bar-Hillel 1993) or that they were simply off-task (Baker et al. 2004). These students should be distinguished from those who invested more efforts, took a trial-and-error approach, but still received the same score of zero.

Time

Consistent with the finding on action sequence length, students who answered the items correctly spent significantly less time to give a response, solving the problem more efficiently than the students who answered incorrectly. On both items, students who solved a problem correctly distributed a significantly larger proportion of their time on last pauses, which is defined as the time elapsed between finalizing one’s answer and the item last appearing on the screen. That is, students who performed better on the item possibly also showed higher self-regulatory skills and tended to allocate more time to metacognitively review their answers for possible errors and reflect on the solution process. Note that the last pause might not fully cover a test-taker’s complete review process and behaviors. For example, if a test-taker noticed an error through reexamination and made a revised D&D action, the pause before the revision was not accounted for in the last pause measure. The shorter last pause for students whose answer was incorrect might be associated with the fact that they were more likely to revise their answers. In addition, reviewing and reflection could occur during problem solving, not necessarily after completing one’s answer, which again was not captured in the last pause measure. It is also possible that a test-taker was simply off-task or bored before proceeding to the next item instead of reviewing. More details are needed (e.g., through cognitive labs or eye-tracking data) in order to understand students’ cognitive and metacognitive processes during the last pause. Furthermore, it is important to note that considering the small effect sizes, the significant differences in the last pause could be related to the large sample size used in the current study.

On the G4 item, students with higher scores showed a significantly shorter first pause. First pause is indicative of the amount of time test-takers spent on processing and encoding the information in the problem stem, building mental representations of the problem, constructing a goal and plans to achieve the goal, performing necessary mental computations to solve the problem, and making decisions on the first D&D action. Therefore, a shorter first pause for the students who scored higher could be related to shorter information processing time taken to comprehend the problem (possibly due to higher literacy proficiency), and/or shorter time to conduct mental computations and decide on the initial step(s). These students, however, spent significantly more time in transitioning between D&D actions, possibly thinking more about the next steps and making decisions on the following solution steps. As mentioned above, considering the very small effect sizes, the significant differences in the time measures might be simply related to the large sample size.

Different results were obtained on the G8 item. On this item, students who successfully solved the problem showed a significantly longer first pause but shorter D&D execution time than those who did not provide a correct solution. This indicated that the students whose answer was correct probably completed the necessary computations (either mentally or through the use of a calculator) and formed a complete answer of all the four digits before starting the execution of the D&D actions, thus taking a longer first pause for planning and shorter dragging-and-dropping time since their answer had been formed. On the other hand, students whose answers were incorrect might have immediately placed a random number into one of the targets or tested a combination of the sources they guessed without systematically planning. As a result, these students took more time to think about the following steps and revise their answers in the execution stage. These results are also consistent with the previous research on expertise that expert problem solvers tended to be better planners and form a better representation of the problems than novices (Chi et al. 1982).

Implications

This research has theoretical implications and shows the value of analyzing the rich process data obtained from interactive mathematics items to infer the complex problem-solving processes and strategies applied by fourth- and eighth-grade students in large-scale educational assessments. While action sequences were analyzed at item level, the process-related measures developed in this study (e.g., strategy classification, answer change patterns, time use) could be generalized across D&D items and could be used to explore other D&D items in future studies. Results suggested that mathematical problem-solving proficiency is related to the acquisition and application of cognitive and metacognitive strategies. Findings also added to the limited previous literature on answer change in digitally-based assessments.

This study's findings may also be of value to educational practice. They could inform test developers' decision-making process when designing digitally-based items. For instance, test developers should be encouraged to use more interactive and technology-enhanced item types, including but not limited to D&D items, to make full use of the process data and better infer mathematical problem-solving processes and skills. Compared to using the "show your work" instruction in traditional paper-and-pencil assessments to elicit the problem-solving procedures, process data on the interactive item types record the detailed steps and decisions test-takers make on each item in a more fine-grained, authentic, and unobtrusive manner. Process data collected from interactive and open-ended items such as simulations could also be used to study constructs that are otherwise difficult to detect and measure such as collaborative problem-solving (Bergner and von Davier 2019) and metacognition (Jiang et al. 2018a, b).

Currently, process data have been mainly considered as a byproduct and the decisions on which information needed to be recorded in process data are not driven by theoretical frameworks or empirical evidences. Similarly, test items are not typically designed with an intention to take full advantage of the process data and identify problem-solving strategies and processes. Given the large amount of information that could be recorded in process data, it is important for assessment developers to plan ahead and pre-define a reasonable number of meaningful events and construct-relevant variables that could provide actionable diagnostic information about test-takers to focus on. For example, the various paths that are representative of common misconceptions or inefficient problem-solving strategies should be identified in the item design stage instead of in post-hoc analysis to identify the low-performing students in real time. Meanwhile, special attention should be paid to reduce the cognitive load and construct-irrelevant variance that might be introduced to these interactive items. For example, Arslan et al. (2020) suggested that surface features such as the physical distance between sources and targets in D&D items could introduce construct-irrelevant variance and should be taken into consideration in item design.

Furthermore, results from this study showed the potential to incorporate process data in scoring rubrics or measurement models to improve test score interpretations and measurement accuracy. Process data used in this study not only provided validity evidence of the item scores, but could be leveraged to enrich the item scores by assigning partial scores or scoring students based on how they arrived at their solutions and the competency of their strategies. For example, test-takers who gave a correct response to the G4 item with three actions could be scored differently from those who exhibited a longer response sequence with many unnecessary steps but still submitted a correct response to indicate the different levels of problem-solving efficiency. In the current study, process data serve an essential role in understanding test-takers' scores on the mathematics items, a Level 3 use of process data based on Bergner and von Davier's (2019) framework. Using process data in the scoring system to infer problem-solving processes could enable us to make higher levels of use (e.g., Level 4 and Level 5) of process data in large-scale assessments, which will in turn illuminate test developers' item design process.

Last but not least, insights about test-takers' cognitive and metacognitive processes and strategies inferred from process data in large-scale assessments could be utilized to provide feedback to teachers and learners who are in need of real-time personalized scaffolding and instruction. That is, they not only could be used for reporting as the summative assessments *of* learning, but also could serve as the assessments *for* learning (Goldhammer et al. 2020). For example, detailed information should be provided to teachers about where the low-performing students struggled or reached an impasse. Such diagnostic information would help educators in identifying the individualized instructional needs of students. Instruction should also be provided to those who received a high score but executed less efficient paths to reinforce their prior knowledge and skills and address their misconceptions to prepare them for future efficient problem solving. It is particularly crucial to provide feedback and instruction on strategy use to students with low prior knowledge (Fyfe et al. 2012; Verschaffel et al. 1999).

Limitations and future research

One of the limitations of this research is that it only focuses on two D&D items administered in NAEP mathematics assessments, one for each grade. Both items assess students' knowledge and skills in numbers and operations. Future analysis could use process data from items on other topics such as geometry and word problems, and items of different difficulty levels to interpret test-takers' cognitive and metacognitive processes on these items and explore the relationships between item-level attributes with these processes. This analysis would also enable us to test the generalizability of the findings obtained from the current study to other topics and items and the developmental stages of mathematical strategies across grades. In addition to the item-level analysis, future research includes extending the exploration of process data to a series of items. For instance, examining the action patterns exhibited by high- and low-performing students across all the items in a test form would reveal the stability and flexibility of test-takers' problem-solving strategies (Elia et al. 2009).

The present research studies the problem-solving strategies and processes by combining the response data with the D&D actions. In addition to the D&D actions, information extracted from other actions (e.g., opening, using, and closing the calculator, opening the scratchpad, calculator keystrokes, scratchwork content, etc.) was also recorded in the format of process data. Incorporating meaningful indices extracted from these relevant actions would provide more comprehensive insights into problem-solving processes. For example, analyzing process data related to calculator use on the G8 item is important for interpreting test-takers' mathematical thinking processes, their approach to solving a problem, and their mathematical and calculator use proficiency. Students could use a calculator as an aid to generate a solution, test a number of possible answers, or check a formed response. Students who came up with a solution using mental computations and used a calculator to check and confirm their answers on the item showed different problem-solving strategies compared to those who used a calculator to test all possible combinations of multiplication they guessed (e.g., $126 \times 7 =$; clear; $127 \times 6 =$; clear; $162 \times 7 =$; clear; $167 \times 2 =$; ...), a sign of gaming the system (Baker et al. 2004). Therefore, future research involves studying whether students used the on-screen calculator and how the calculator was used (Jiang and Cayton-Hodges, n.d., under review). Similarly, studying student use of the digital scratchpad and the scratchwork created on items such as the G4 item enables us to understand how students visualized mental representations and solved problems.

Results from NAEP have documented achievement gaps in mathematics (Plucker et al. 2010). Closing achievement gaps is an important topic in education policy and research (Flores 2007). Individual differences in academic achievement and performance could sometimes be understood through differences in problem-solving strategies and processes (He et al. 2019). To this end, process data provide unique insights into achievement gaps between student subgroups. For example, with process data we will understand whether males and females tend to apply different strategies when attempting to solve problems, which would further shed light on the gender differences in mathematical problem-solving processes and outcomes. Therefore, future research should include subgroup analysis of the measures developed from process data to understand the gaps in cognitive and metacognitive processes between various population subgroups.

Abbreviations

NAEP: National Assessment of Educational Progress; DBA: Digitally-based assessments; D&D: Drag and drop; G4 item: Grade-Four item; G8 item: Grade-Eight item; SRL: Self-regulated learning; K-12: From kindergarten to 12th grade.

Acknowledgements

We would like to thank Sooyeon Kim, Mary Louise Lennon, and Ann Kennedy for their insightful feedback on the manuscript.

Authors' contributions

YJ and LES developed the idea for the study. YJ and TG cleaned the data. YJ wrote the first version of the manuscript and conducted all the analyses. All authors discussed the results and contributed to the final version of the manuscript. All authors read and approved the final manuscript.

Funding

This project has been funded at least in part with Federal funds from the U.S. Department of Education under contract number 99190019C0045. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available. Data are however available from the National Center for Education Statistics for restricted use upon reasonable request.

Competing interests

All authors declare that they have no competing interests.

Received: 4 June 2020 Accepted: 9 February 2021
Published online: 24 February 2021

References

Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing, 22*(4), 509–531. https://doi.org/10.1191/0265532205lt317oa.

Arslan, B., Jiang, Y., Keehner, M., Gong, T., & Katz, I. R. (2020). The Effect of Drag-and-Drop Item Features on Test-Taker Performance and Response Strategies. *Educational Measurement: Issues and Practice, 39*(2), 96–106. https://doi.org/10.1111/emip.12326.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the Cognitive Tutor classroom: When students "game the system." *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383–390.

Bauer, D., Kopp, V., & Fischer, M. R. (2007). Answer changing in multiple choice assessment change that answer when in doubt - and spread the word! *BMC Medical Education, 7*(28), 1–5. https://doi.org/10.1186/1472-6920-7-28.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological), 57*(1), 289–300.

Bergner, Y., & von Davier, A. A. (2019). Process Data in NAEP: Past, Present, and Future. *Journal of Educational and Behavioral Statistics, 44*(6), 706–732. https://doi.org/10.3102/1076998618784700.

Bryant, W. (2017). Developing a Strategy for Using Technology-Enhanced Items in Large-Scale Standardized Tests. *Practical Assessment, Research & Evaluation*, *22*(1), 1–10. Retrieved from https://pareonline.net/getvn.asp?v=22&n=1

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30*(4), 277–291.

Cai, J., & Cifarelli, V. (2005). Exploring mathematical exploration: How two college students formulated and solved their own mathematical problems. *Focus on Learning Problems in Mathematics, 27*(3), 43.

Cai, J., Silber, S., Hwang, S., Nie, B., Moyer, J. C., & Wang, N. (2014). Problem-solving strategies as a measure of longitudinal curricular effects on student learning. *Proceedings of the Joint Meeting 2 - 73 of PME 38 and PME-NA 36*, *2*, 233–240.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence* (Vol. 1, pp. 7–76). Hillsdale, NJ: Erlbaum.

Elia, I., van den Heuvel-Panhuizen, M., & Kolovou, A. (2009). Exploring strategy use and strategy flexibility in non-routine problem solving by primary school high achievers in mathematics. *ZDM - International Journal on Mathematics Education, 41*(5), 605–618. https://doi.org/10.1007/s11858-009-0184-6.

Flores, A. (2007). Examining Disparities in Mathematics Education: Achievement Gap or Opportunity Gap? *The High School Journal, 91*(1), 29–42. https://doi.org/10.1353/hsj.2007.0022.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18. https://doi.org/10.1037/a0024338.

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology, 104*(4), 1094–1108. https://doi.org/10.1037/a0028389.

Galbraith, D., & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In K. P. H. Lindgren, Eva; Sullivan (Ed.), *Observing writing: Insights from keystroke logging and handwriting* (pp. 306–325). Brill.

Goldhammer, F., Scherer, R., & Greiff, S. (2020). Editorial: Advancements in technology-based assessment: Emerging item formats, test designs, and data sources. *Frontiers in Psychology, 10,* 1–4. https://doi.org/10.3389/fpsyg.2019.03047.

Gong, T., Shuai, L., Arslan, B., & Jiang, Y. (2020). Process based analysis on scientific inquiry tasks using large-scale national assessment dataset. In *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020* (pp. 417–423).

Goos, M. (2002). Understanding metacognitive failure. *The Journal of Mathematical Behavior, 21*(3), 283–302.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173–183. https://doi.org/10.1080/08957347.2016.1171766.

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*(November), 1–15. https://doi.org/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.

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*. (October). https://doi.org/https://doi.org/10.1787/650918f2-en

Hoyles, C., & Noss, R. (2003). What can digital technologies take from and bring to research in mathematics education? *Second international Handbook of Mathematics Education* (pp. 323–349). Dordrecht: Springer.

Jiang, Y., & Cayton-Hodges, G. A. (n.d.). *Student use of calculators during mathematical problem solving in a large-scale digitally-based assessment*.

Jiang, Y., Clarke-Midura, J., Baker, R. S., Paquette, L., & Keller, B. (2018). How immersive virtual environments foster self-regulated learning. In R. Zheng (Ed.), *Digital Technologies and Instructional Design for Personalized Learning* (pp. 28–54). https://doi.org/https://doi.org/10.4018/978-1-5225-3940-7.ch002

Jiang, Y., Clarke-Midura, J., Keller, B., Baker, R. S., Paquette, L., & Ocumpaugh, J. (2018). Note-taking and science inquiry in an open-ended learning environment. *Contemporary Educational Psychology, 55,* 12–29. https://doi.org/10.1016/j.cedpsych.2018.08.004.

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 61–78). Cambridge, MA: Cambridge University Press.

Kramarski, B., Mevarech, Z. R., & Arami, M. (2002). The effects of metacognitive instruction on solving mathematical authentic tasks. *Educational Studies in Mathematics, 49*(2), 225–250.

Lee, Y., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education, 2*(8), 1–24.

Lester, F. K. (1994). Musings about mathematical problem-solving research : 1970–1994. *Journal for Research in Mathematics Education*, *25*(6), 660–675. Retrieved from http://www.jstor.org/stable/749578

Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of Response Changes in the GRE Revised General Test. *Educational and Psychological Measurement, 75*(6), 1002–1020. https://doi.org/10.1177/0013164415573988.

Mcconnell, M. M., Regehr, G., Wood, T. J., & Eva, K. W. (2012). Self-monitoring and its relationship to medical knowledge. *Advances in Health Sciences Education, 17*(3), 311–323. https://doi.org/10.1007/s10459-011-9305-4.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design (Research Report 03–16)*. Princeton, NJ.

Montague, M., & Bos, C. S. (1990). Cognitive and metacognitive characteristics of eighth grade students' mathematical problem solving. *Learning and Individual Differences, 2*(3), 371–388.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. https://doi.org/https://doi.org/10.1111/j.1949-8594.2001.tb17957.x

Ni, Y., & Zhou, Y. D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist, 40*(1), 27–52. https://doi.org/10.1207/s15326985ep4001_3.

Özsoy, G., & Ataman, A. (2009). The effect of metacognitive strategy training on mathematical problem solving achievement. *International Electronic Journal of Elementary Education, 1*(2), 68–82.

Pape, S. J., & Wang, C. (2003). Middle school children's strategic behavior: Classification and relation to academic achievement and mathematical problem solving. *Instructional Science, 31*(6), 419–449. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2003-08139-004&site=ehost-live pape.12@osu.edu

Plucker, J., Burroughts, N., & Song, R. (2010). The growing excellence gap in K-12 education: Mind the (Other) Gap! In *Center for Evaluation and Education Policy*. Retrieved from http://www.jkcf.org/assets/1/7/ExcellenceGapBrief_-_Plucker.pdf

Polya, G. (1957). *How to solve it* (2nd ed.). Princeton, NJ: Lawrence Erlbaum.

Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education, 9*(1), 1–17. https://doi.org/10.1186/s40536-020-00092-z.

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education, 20*(1), 8–27.

Roche, A. (2010). Decimats: Helping students to make sense of decimal place value. *Australian Primary Mathematics Classroom, 15*(2), 4–12.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project (NCES 2005–457)*. Washington, DC.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in E-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6), 3–44.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334–370). New York, NY: MacMillan.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 293–312.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*(1), 5–20. https://doi.org/10.3102/1076998607302626.

Verschaffel, L., Corte, E. D., Lasure, S., Vaerenbergh, G. V., Bogaerts, H., & Ratinckx, E. (1999). Understanding How Students Develop Mathematical Models. *Mathematical Thinking and Learning, 1*(3), 195–229. https://doi.org/10.1207/s15327833mtl0103.

Westenskow, A., Moyer-Packenham, P. S., Anderson-Pence, K. L., Shumway, J. F., & Jordan, K. (2014). Cute drawings? What students' fractional representations reveal about their whole number bias. In *Proceedings of the 12th International Conference of the Mathematics Education into the 21st Century Project* (pp. 1–6).

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Hillsdale, NJ: Lawrence Erlbaum Associates.

Yimer, A., & Ellerton, N. F. (2010). A five-phase model for mathematical problem solving: Identifying synergies in pre-service-teachers' metacognitive and cognitive actions. *ZDM - International Journal on Mathematics Education, 42*(2), 245–261. https://doi.org/10.1007/s11858-009-0223-3.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337–362. https://doi.org/10.1207/S15324818AME1504.

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.