Process based Analysis on Scientific Inquiry Tasks Using Large-scale National Assessment Dataset

Tao Gong, Lan Shuai,
Burcu Arslan, Yang Jiang
Educational Testing Service
tgong@ets.org

ABSTRACT

This paper investigates differences in students having various scores when designing controlled experiments in two types of scientific inquiry tasks (a fair test and an exhaustive test). We measure temporal features of preparation time and execution time, which reflect respectively the process of question understanding and answer planning and that of executing the control-of-variables strategy in answer formulation. We also measure mean execution time per answering event to reflect the efficiency of answering events. Results show that: in the fair test, the full score students showed less execution time than the lowest score ones; in the exhaustive test, the full score students showed more execution time than the lowest score ones; but in both tests, the high-performing students had less mean execution time than the low-performing ones. These results reveal that despite test differences, students who appropriately apply the control-of-variables strategy in these tests are more goal-directed and efficient in planning and executing response strategies than those who fail to do so. This study provides process-based features and large-scale evidence of scientific inquiry practice in in educational assessment.

Keywords

Control-of-variables strategy, preparation time, execution time

1. INTRODUCTION

Scientific inquiry refers to the activities by which students develop knowledge of scientific ideas and understand how to investigate the natural world in a scientific way [1]. In STEM education, scientific inquiry skills have been emphasized as a key goal of scientific literacy [2,3], and scientists and science educators have advocated teaching science as inquiry [4–8]. Among scientific inquiry activities (see [6] for overview), *planning*, *designing*, and *carrying out investigations* have long become a principal focus of children's and youngsters' scientific inquiry practices [9,10]. Many studies aim to investigate, based primarily on response data, how students design controlled experiments by constructing related conditions for comparison.

Fair tests and exhaustive tests have been widely adopted to examine how students plan, design, and carry out controlled experiments. A *fair test* (see an example in Sec. 2.2) refers to a controlled investigation carried out to answer a scientific question about the effect of a target variable. To control for confounding factors and be scientifically sound, students are supposed to apply a *control-of-variables strategy* (CVS) [9,11] to ensure that: (a) all the other variable(s) are kept constant; and (b) only the variable(s) under investigation is changed across conditions for comparison. Only in such a fair setting, the effect of the target variable(s) can be explicitly observed, since the other variables remain constant across conditions. Students can complete the task by choosing, among a large number of possible combinations of variables, one or a few conditions that meet the fair test requirement.

An *exhaustive test* (a.k.a. combinatorial test, [12,13]) (see an example in Sec. 2.3) requires constructing, physically or mentally, all possible combinations of given variables to address inquiry on which conditions could cause a specific outcome. Like fair tests, students in exhaustive tests also need to control target variables to construct combinations, but the number of possible combinations is generally smaller than that in fair tests. In exhaustive tests, students are asked to enumerate all combinations; in fair tests, students only need to select one (or a few) condition that meet the requirement. In this sense, exhaustive tests require more cognitive resources especially in situations with not easily foreseen combinations. How to conduct an exhaustive test is taught and learned late in science education, and items assessing such skill often lie in the 8th, 12th, or higher-grade assessments [3].

CVS is required in both types of tests. Among other types of procedural knowledge, or "process skills", CVS is deemed central to early science instruction [14]. Existing research shows that children, adolescents, and adults with low scientific inquiry expertise tend to have difficulty in applying CVS [9,10,15]. However, due to lacking measures on processes of scientific inquiry, existing studies focus primarily on students' responses.

In modern digitally-based assessment programs (e.g., National Assessment of Educational Progress (NAEP)), technology-enhanced (TE) items have been used to study scientific inquiry practice. The interactive nature of such items allows recording not only final submitted answers, but also the process whereby students formulate their answers via a series of drag-and-drop, (de)selection, or correction actions. Obtained process data can gather additional evidence on what students do during inquiry [16–18]. TE items have now touched upon many disciplines, including math, science, and social science [18–21], and process data obtained have covered not only observable behaviors of test-takers in problem solving but also frequencies and durations of such actions, both contributing to illustrating the mastery phases in scientific inquiry and response strategies of students [22–25]. In addition, process-based analyses help discover the aspects where students of different scores differ, and lead to better understanding of the cognitive framework of scientific inquiry.

Tao Gong, Lan Shuai, Burcu Arslan and Yang Jiang "Using Process Data to Evaluate Scientific Inquiry Practice in Technology-Enhanced Assessment" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 417 - 423

Rather than concrete events recorded in obtained process data of TE items, the time needed for different stages of scientific inquiry has been undervalued in recent research of scientific inquiry or problem solving [26]. Temporal information can reveal different stages of problem solving, clarify performance patterns of students with different levels of problem solving competency, and allow inferring something about the cognitive processes occurring at various phases of problem solving.

Noting these, this study aims to investigate the scientific inquiry practice, to be specific, the practice of designing controlled experiments by applying the CVS in fair and exhaustive tests. By evaluating relations between defined process-based, temporal measures and students' performance gauged by scores, we aim to address the following two research questions:

(a) What are the process-based characteristics of the high-/low-performing (indicated by different scores) students in the tests?

(b) Are these process-based characteristics consistent across the fair and exhaustive tests?

Answers to these questions can benefit the general discussions on scientific inquiry practice, especially whether the CSV strategy manifest differently across various types of inquiry tasks. They also provide actionable feedback to teaching and learning the skills required in scientific inquiry tasks. Moreover, this study enriches the literature of using process data and derived features to address theoretical issues in educational assessment.

In the rest of the paper, we describe the NAEP science fair test and exhaustive test used in this study, define the process-based measures, and describe the analysis plan. Then, we report the results, discuss the research questions accordingly, and conclude the study by highlighting theoretical or operational applications of process-based analyses in education and psychology research.

2. METHORDS AND MATERIALS

2.1 NAEP Science Tasks

Our study uses the 2018 NAEP science pilot tasks. NAEP is a congressionally mandated, nationwide digital assessment project administered by the National Center for Education Statistics (NCES) in the Institute of Education Sciences of the U.S. Department of Education. NAEP provides large-scale, regular assessments on many disciplines (e.g., math, reading, writing, science, etc.). All the assessments are designed and updated by content specialists, education experts, and teachers from around the U.S. Participants of the tests are grades 4 (~9-year-olds), 8 (~13-year-olds) and 12 (~17-year-olds) students. Along with the assessment, survey data of students, teachers, and schools are gathered, covering students' demographical information (gender and ethnicity), special programs, self-evaluation of performance, etc. NAEP has now become one of the largest and most important national assessments of what U.S. students know and can do.

The 2018 assessment was conducted by the NAEP field staff, who went into schools across the nation to administer tasks on students from the NAEP sample. The science tasks were administered on NAEP-provided tablets with an attached keyboard and earbuds. Students had 60 minutes to complete the questions in the given task. Tutorials and surveys were given throughout the test.

A total of 32 science tasks were designed for the 2018 NAEP pilot test, some of which were administered on grades 4, 8, and 12 students. Our study focuses on a fair test and an exhaustive test, which were administered respectively on grade 8 and 12 students. This choice was due to three considerations. First, lower grade students have not been taught how to solve both types of tests, so we avoid tasks administered on grade 4 students. Second, since fair tests were administered mostly on grade 4 and 8 students but exhaustive tests were administered mostly on grade 12 students, we could not select fair tests and exhaustive tests administered on students of the same grade. Third, to properly answer the two chosen tests, students needed to submit similar numbers of distinct answers, which avoided possible interference from cognitive load in students' answer formulation process.

Due to the privacy and secure nature of the NAEP data, we use conceptually equivalent tasks (*cover tasks*) to disguise the content and context of the real tasks. Cover tasks have similar underlying structures and require similar cognitive processes to solve.

2.2 Fair Test, Scoring Rubric, Students

This test came from an earth and space science task. Its cover test is as follows (see Figure 1). A city near a mountain suffers from north winds each year. Its government plans to test the wind-blocking power of three types of trees, which can be planted at the north side of the mountain. After simple instructions of the task, in the fair test scene of the task, students are asked to drag each type of trees and drop them at one of the four virtual mountains resembling the real one near the city. Students can drop the trees at the foot (low), side (medium), or peak (high) of the north side of a mountain. Each mountain holds at most one type of trees, and each type can only be planted at one mount. Students can move trees from one position/mountain to another. After selecting the locations of the three types of trees, students can click on the on-screen "Submit" button to trigger the experiment, and the wind speeds before and after passing over the mountains are shown on the screen. By default, one mountain is left without any tree.

There are two types of variables in this fair test: tree type and tree position on mountain. To illustrate the effect of trees, students must control the positions of the trees to be identical across conditions (mountains). There are in principle 3×3×3×$P(4,3)$ = 648 choices for students to plant the trees, among which 3×$P(4,3)$ = 72 choices meet the fair test requirement.

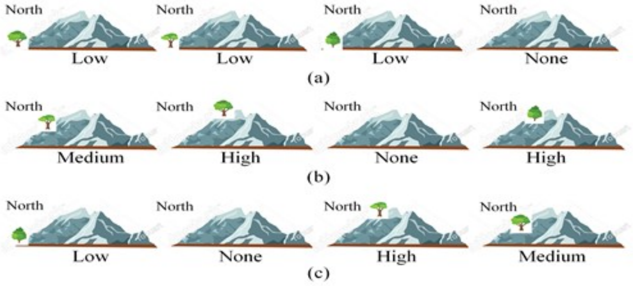

Figure 1. Example answers in the fair test. "Low", "Medium", "High" denote positions (foot, side, peak) of trees in the north side of the mountain. "None" means no tree planted. In (a), the first "Low" indicates that one type of trees are planted at the foot of the mountain, the second and third "Low" indicate that the other types of trees are planted on the second and third mountains, and "None" means no trees on the fourth mountain. The scoring rubric ignores tree types and the mountain without trees, the submitted answer can thus be denoted by the positions of trees in three mountains.

Table 1 shows the scoring rubric of this test. The rubric ignores tree type, since students cannot put the same type of trees in two mountains or two positions of one mountain. It also ignores the mountain without trees ("None"), since this is a default condition of the test; no matter how to answer the test, one mountain must be left without trees. A complete comparison to show the effects of trees requires the condition without trees, but in this test, students are not required to set up this condition. The rubric focuses on the target variable of tree positions across mountains. Answers meeting the fair test requirement receive a full score (3), those partially meeting the requirement get a partial score (2), and those not meeting the requirement have the lowest score (1).

Table 1. Scoring rubric of the fair test.

Score	Rubric
3	Choices of trees have the same positions on three mountains (e.g., Low; Low; Low in Figure 1(a))
2	Two types of trees are on the same positions of mountains (e.g., Medium; High; High in Figure 1(b))
1	Positions of the three types of trees on mountains are all distinct (e.g., Low; High; Medium in Figure 1(c))

This task was administered to 1,657 (825 females) grade 8 students. The response and process data of 1,607 (800 females) students were recorded in the fair test for analyses. Fifty-one students, due to various reasons, quit before reaching the fair test.

2.3 Exhaustive Test, Scoring Rubric, Students

This test came from a life science task. Its cover test is as follows. Farmers attempt to cultivate flowers with a special color in a natural way (without using any fertilizers) or using two types of fertilizers. After simple instructions of the task, students are asked to design an experiment to show which way has the highest probability to induce the target color. They can set up a condition by selecting (or not) any (or both) type of the fertilizers. After setting up a condition, they can click on the on-screen "Save" button to save the condition. They can also click on a saved condition and click on the "Delete" button to remove it. After setting up and saving many conditions, students can click on the "Submit" button to submit saved conditions as final answers.

This is a typical exhaustive test with four possible combinations of the variables (see Figure 2). The conditions no fertilizer (Figure 2(a)) and both fertilizers (Figure 2(d)) are not easily foreseen.

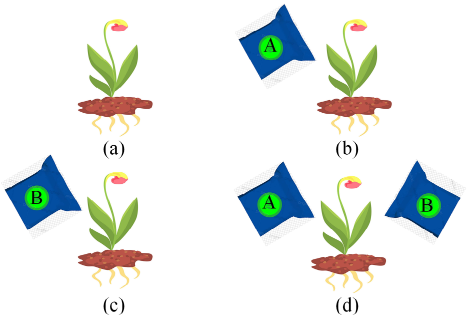

Figure 2. All combinations in the exhaustive test: (a) None; (b): A; (c): B; (d): A + B.

Table 2 shows the scoring rubric of the test. It has four scales, among which partially high (3) and partially low (2) are classified by submitted saved conditions, especially whether they include the not-easily foreseen conditions. Whether the rubric reasonably classifies students' skill levels is not the focus of this paper.

This task was administered to 2,869 (1,360 females) grade 12 students. The response and process data of 2,726 (1,285 females) students were recorded in the exhaustive test for the analyses. Due to various reasons (e.g., early quitting or glitches in data capture), the process data of 173 students were missing.

Table 2. Scoring rubric of the exhaustive test.

Score	Rubric
4	Saved conditions cover all four conditions in Figure 2
3	Saved conditions do not include the condition of Figure 2(a), OR do not include the condition of Figure 2(b) or Figure 2(c)
2	Saved conditions do not include the condition of Figure 2(d), OR do not include the conditions of Figure 2(b) or Figure 2(c), OR do not include both the conditions of Figure 2(a) and Figure 2(d)
1	Saved conditions do not match the above cases

2.4 Process-Based Measures

The NAEP digital assessment system can recorded students' process data in these interactive TE items. Such data consisted of a list of activity logs plus their time stamps. Activities included user events (e.g., drag-and-drop, save, delete, or correct, etc.) and system events (e.g., play instructions or video clips). They allow reconstructing submitted answers, tracing sequences of students' drag-and-drop or saving/deletion/correction actions, and durations of these activities. Based on such data, we propose and measure three temporal measures, namely preparation time, execution time, and mean execution time per answering event.

Preparation time (PT) is defined as the duration between students enter the test scene and make their first answer-related event, such as drag-and-drop one type of trees, select a fertilizer, or save a condition without any fertilizers. Before the test scenes, students were given instructions and practice trials on how to set up answers in the test scenes. Therefore, PT does not involve the time students spent on getting familiar with the system. PT reflects the time for students to read and understand instructions, as well as think and get ready to formulate their answers.

Execution time (ET) is defined as the duration between students' first and last answer-related events. The ending time point of ET was not when students clicked on the submission button. This is because we do not know exactly whether students reviewed their answers after making their last drag-and-drop or selection event before submission. If they did review and made corrections, the measure can certainly capture such reviewing event; if they did not make any changes, it is unclear whether the time between the last answer-related event and the submission event was spent on reviewing. Many students actually clicked on the "Submit" button immediately after the last answer-related event.

ET is the sum of the durations of different numbers of answer related events. In the fair test, such events include dragging and dropping a type of trees to a mountain or moving one from one mountain to another; in the exhaustive test, such events include selecting one or two fertilizers, or saving a condition or cancelling a saved one. Students having different performances may put different efforts when conducting these events, and different tasks may require different numbers of events to formulate answers, which already lead to different ET. Noting these, we also calculate the *mean execution time per answering event* (MET). MET is operationalized as the execution time divided by the number of answering events. ET reflects the total efforts made by students to construct answers, including setting up, revising or (possibly) reviewing their choices, whereas MET reflects the average effort made to construct their answers, and it controls the effect induced by different numbers of events.

Apart from temporal measures, one can also measure the numbers of answer related events made by students during the answering process. However, for students who conducted the same number of answering events, this count-based measure cannot clarify how much effort each event costs to these students; more events may not always require more efforts, since an efficient test-taker can conduct many events in a short period of time; and more events alone cannot predict performance in different tests, since some of the events could be answer revisions, which simply indicate low efficiency. The temporal measures defined in our study avoid these confusions and are more informative of students' degrees of efficiency in designing controlled experiments in those tasks.

2.5 Analyses

For each dataset, we take a 98% winsorization estimation [27] to remove spurious outliers. We also remove the missing values.

We conduct two types of analyses. First, we check how many students appropriately applied the required CVS in the tests based on score distributions and illustrate the frequent (top 10) correct or incorrect submitted answers. Second, treating score as a ranked variable, we conduct the Kruskal-Wallis test [28], a non-parametric version of ANOVA test, to compare students' scores and the three measures across score groups. If the omnibus test produces a significant p-value, we conduct the Wilcoxon signed-rank test on pair-wised score groups to clarify which two groups have different population means of the measures. This test is also non-parametric. These two statistical tests provide direct evidence on the relation between students' performance (scores) and the process-based measures. The tests are implemented using the kruskal.test and wilcox.test functions in the *stats* package in R 3.6.1 [29]. For each task, there are three Kruskal-Wallis tests respectively on three measures, accordingly, the critical p value for identifying significance is set to $.05/3 \approx .0167$.

3. RESULTS

3.1 Fair Test

In this test, 41.4% of the students had the lowest score (1), and only 29.5% properly applied the CVS and got the full score (3). The rest (29.1%) received a partial score (2).

Figure 3 shows the top 10 frequent answers submitted by students. It shows that "Low; Low; Low" is the most frequent correct answer, but other correct answers like "Medium; Medium; Medium" and "High; High; High" are less so. In addition, "Low; Medium; High" is the most common wrong answer. Its variations, such as "High; Medium; Low" or "Low; High; Medium", are also frequent, but all of them receive the lowest score (1) (see Table 1). Answers receiving a partial score (2) (e.g., "Medium; Low; Medium") are less frequent, compared to other types of answers. These results indicate that over 70% of students did not properly apply the CVS strategy in this scientific inquiry task.

Table 3 shows the means and standard errors of the process-based measures in each score group. The Kruskal-Wallis tests report significant differences in PT ($\chi^2 = 12.2$, df = 2, $p = .002$), ET ($\chi^2 = 89.916$, df = 2, $p < .001$), and MET ($\chi^2 = 64.776$, df = 2, $p < .001$) between score groups. Table 4 shows the Wilcoxon signed-rank tests results. It reveals that the full score students had significantly shorter PTs, ETs, and METs than the lowest and partial score students, but these measures were not significantly different between the lowest and partial score students.

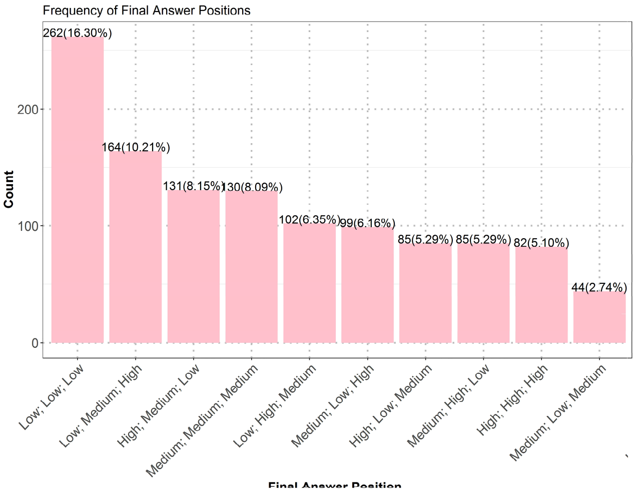

Figure 3. Top 10 frequent answers in the fair test. Numbers on top of bars are numbers of students and those inside brackets are proportions of students.

Table 3. PT, ET and MET across score groups. Numbers (in seconds) outside brackets are means and those inside are standard errors.

Score	PT	ET	MET
1	85.571 (1.166)	41.330 (1.098)	5.125 (.091)
2	85.154 (1.407)	38.807 (1.216)	4.958 (.103)
3	79.745 (1.172)	29.082 (1.081)	4.138 (.090)

Table 4. Wilcoxon signed-rank test results in the fair test. "1" to "3" in the first column denote score groups. Values outside brackets are test statistics, and those inside are *p* values. Significant results are marked in bold.

	PT	ET	MET
1v2	158942 (.527)	**163023 (.016)**	158766 (.548)
1v3	**176639 (.001)**	**2038350.5 (.001)**	**199945.5 (.001)**
2v3	**120966.5 (.014)**	**139637.5 (.001)**	**136592.5 (.001)**

3.2 Exhaustive Test

In this test, 25.2% of the students received the lowest score (1), and 33.9% properly applied the CVS strategy and received the full score (4). The rest received the partially high (3) (34.1%) and partially low (2) (6.8%) scores.

Figure 4 shows the top 10 frequent answers, among which "A; B; A + B; None" and its variations "A; A + B; B; None" and "A + B; A; B; None" receive the full score (4), but they are not frequent compared to the answers "A + B", "B", "A", and "None", which are among the most frequent answers and receive the lowest score (1) (see Table 2). The answers having partially high (e.g., "A; A + B; None") or low (e.g., "A; A + B") scores are less frequent. These results show that many students did not have the required scientific inquiry skill.

Table 5 shows the means and standard errors of the process-based measures in each score group. The Kruskal-Wallis test report significant differences in PT (χ^2 = 127.69, df = 3, p < .001), ET (χ^2= 332.88, df = 3, p < .001), and MET (χ^2 = 238.93, df = 3, p < .001) between the score groups. Table 6 shows the Wilcoxon signed-rank tests results. It reveals that the lowest score students had significantly longer PTs than the students from other score groups. Unlike the fair tests, the lowest score students had significantly shorter ETs than the full and partial score students. Like the fair tests, the lowest score students had significantly longer METs than the full score students.

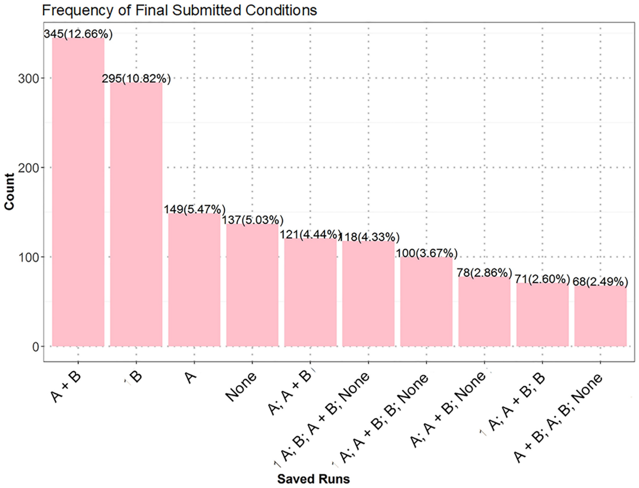

Figure 4. Top 10 frequent answers in the exhaustive test. Numbers on top of bars are numbers of students and those inside brackets are proportions of students.

Table 5. PT, ET, and MET across score groups. Numbers (in seconds) outside brackets are means and those inside are standard errors.

Score	PT	ET	MET
1	9.056 (.325)	24.949 (.922)	5.502 (.144)
2	6.797 (.439)	41.623 (1.804)	3.520 (.113)
3	7.105 (.207)	31.700 (.715)	3.899 (.070)
4	5.714 (.172）	42.523 (.763)	3.140 (.051)

Table 6. Wilcoxon signed-rank test results in the exhaustive test. "1" to "4" in the first column denote score groups. Values outside brackets are the test statistics, and those inside are *p* values. Significant results are marked in bold.

	PT	ET	MET
1v2	**75018.0 (< .001)**	**29065 (< .001)**	**83948 (< .001)**
1v3	**372475.5 (< .001)**	**215673.5 (< .001)**	**400288.5 (< .001)**
1v4	**422941.5 (< .001)**	**128978.5 (< .001)**	**458813.5 (< .001)**
2v3	84443.5 (.693)	**11656 (< .001)**	80219.5 (.147)
2v4	**98433.5 (< .005)**	81207 (.284)	**100851 (< .001)**
3v4	**501936.5 (< .001)**	**274362.5 (< .001)**	**531023.5 (< .001)**

4. DISCUSSIONS

Based on two NAEP science tasks (a fair test and an exhaustive test) and three process-based temporal features, we dig out, from both response and process data, the differences and similarities between the high-/low-performing students in those two typical types of scientific inquiry practice.

As for response, the score distributions illustrate that many (over 70%) grade 8 or 12 students failed to properly apply the control-of-variables strategy in the fair and exhaustive tests, consistent with the previous literature [9]. In addition, in the fair test (see Figure 3), the most commonly wrong strategy is to vary both variables' levels at the same time, e.g., "Low; Medium; High" and its variations. This is also shown in previous observations [17]. In the exhaustive test (see Figure 4), the most commonly wrong strategy is to save only one of the four possible conditions as in Figure 2. This indicates that the low-performing students probably did not have the intention or the capability to design a controlled experiment but simply guessed an answer.

As for process, rather than specific actions or sequences of drag-and-drop actions as in recent studies on TE items [30], our study defines temporal features and adopts non-parametric statistical tests on these stage-level features to reveal quantitative differences between the high- and low-performing students.

The statistical tests collectively show that: in terms of preparation, the full score students spent less time before making their first answering related activity in both the fair and exhaustive tests, which are consistent with other studies [30]. Longer preparation time in the lowest score students shows that such low-performing students might have difficulty in quickly grasping the instructions or need more time to think before taking any action, whereas the high-performing students could efficiently grasp the instructions and foresee the required conditions. These results suggest that the different performances between the full and lowest score students have already manifested at the early stage of scientific inquiry practice, where no answer is formulated. In other words, whether a student can appropriately apply the control-of-variable strategy in a fair task could be highly correlated with whether he or she can efficiently grasp the instruction at the beginning of the task.

In terms of execution time, there exist differences between the fair and exhaustive tests. In the fair test, the lowest score students spent longer time on conducting the drag-and-drop actions to construct answers. As shown in Figure 3, their submitted answers after such a long execution time still failed to meet the fair test requirements. This echoes the fact that these students did not follow the instructions nor get well prepared for the fair tests. To be specific, in the fair test, the minimum number of events to construct an answer was three (dragging and dropping each type of trees respectively to the same or different locations of three mountains). Two possible situations lead to longer execution time in the lowest score students: they conducted many revisions to their early choices, or spent more time on conducting each activity, indicating their hesitation or uncertainty about their choices, or more time needed to come up with a solution due to a lack of relevant domain knowledge. Here, the results of mean execution time per answering event (see Table 4) reveal that no matter how many revisions they conducted, on average, the lowest score students spent more time on setting up each of their answers than the full score students; i.e., the full score students were more efficient than others.

In the exhaustive test, constructing all possible conditions is not trivial and requires more resources and related events. As shown in Table 5, the lowest score students spent shorter time in constructing or revising their saved conditions, whereas the full score students spent longer time in doing so. As in Figure 4, the lowest score students (and those having partial scores) did not save enough conditions, but the full score students submitted each of the possible conditions as required by the test. Therefore, the longer execution time of the full score students reflects the fact that these high-performing students had endeavored to set up all required conditions before the final submission. By contrast, the shorter execution time of the lowest score (and partial score) students indicates that: (a) these low-performing students did not spend much time on exploring possible conditions but completed the test by submitting lack-of-thinking results, indicating their low motivation or lack of engagement in problem solving; or (b) throughout the test, they might not realize that they needed to save and submit all possible conditions, so they simply submitted one condition and left the test. Both cases are consistent with the response data of frequent wrong answers submitted (see Figure 4), but they point to different causes of failing the test.

Since the numbers of conditions saved are different across score groups, comparing the execution time, which is the sum of the duration of different numbers of actions, is not enough to reflect whether the efficiency of high- or low-performing students is similar. We need to further compare the mean execution time per answering event. The full score students spent less time (see Table 6) on conducting each answering related action than the low-performing students. This implies that although the full score students conducted more actions, they were more efficient, by putting less effort on each action, than the lowest score students (and those having partial scores). In this sense, the results in the two tests are consistent: the students who properly apply the control-of-variable strategies show more goal-directed and efficient behaviors [30] than those who failed to do so.

The contrasting results of execution time between the fair and exhaustive tests reveal the differences between the two types of scientific inquiry practice. Although both tests require controlling variables under investigation, the nature of control is different, so are the required cognitive resources to properly complete the tests. In the fair test, to study the effect of a target variable (tree type, see Figure 1), students need to keep the other variable (tree position) unchanged. In the exhaustive test, students need to combine different values (use or not use, see Figure 2) of the variables (fertilizers A and B) to set up a set of conditions for comparison. Properly completing this test requires mentally constructing the conditions not easily foreseen and spending time and energy in thinking and setting up each possible condition, thus requiring more cognitive resources than the fair test, the latter of which only requires adjusting the target variable and holding the other one(s) constant. These results indicate that the same control-of-variables strategy manifests differently in different scientific inquiry practices. Systematic teaching and learning of this strategy require task-specific training in different situations.

All the results reveal the aspects in which high-performing students excel low-performing ones, including: (a) grasping instructions, (b) extracting requirements, and (c) constructing answers. Compared to high-performing students, low-performing students had lower efficiency in grasping necessary knowledge and applying required strategies in the tests. As a consequence, in the fair test, low-performing students struggled in selecting and revising answers, and ended up submitting wrong answers; and in the exhaustive test, they failed to envision all possible conditions, and failed to construct enough conditions as the final answers.

The above discussions focus primarily on statistical differences between the full and lowest score students. This is because that our statistical analyses report consistent results between the two score groups. However, results are not consistent when partial score groups are involved. Such inconsistency could be due to several reasons. First, some partial score groups contained fewer students than the other two groups. Second, according to the scoring rubrics, the response difference between the full (or the lowest) score and a partial score is smaller than that between the full and lowest scores, which may cause smaller difference in answering events and/or their durations. Both factors reduced the statistical power of the analyses. Third, lacking empirical basis, the predefined score rubrics might not clearly differentiate students having different levels of problem solving competency. This issue is beyond the scope of the current study. Nonetheless, such inconsistency calls for statistically more powerful process-based features to reveal the differences between students having good and poor performances in science inquiry practice and understand how they apply required skills in such practice.

5. CONCLUSIONS

This study makes use of three process-based, temporal measures to analyze how students conduct scientific inquiry in practice. We identify both the global (e.g., durations of thinking, and total duration of execution) differences and local (e.g., execution efficiency) consistency between students who can appropriately apply the control-of-variables strategies in scientific inquiry practice and those who fail to do so. The findings provide new evidence to the general discussions of the relations among individual capacity (e.g., control-of-variables strategy), nature of test (e.g., fair or exhaustive test), problem-solving process (e.g., duration and efficiency of activities), and assessment performance (e.g., submitted answers and scores). The process-based features have proven values in revealing performance differences in the fair and exhaustive tests. Analysis results based on these measures reveal the aspects or stages during the problem-solving process in which teachers can provide guidance or students can self-improve to teach the required inquiry skills or properly apply them, thus improving students' performances in science inquiry practice.

6. ACKNOWLEDGMENTS

This work was supported in part by National Center for Education Statistics (NCES) and Educational Testing Service (ETS). We thank Kathleen Scalise, Madeleine Keehner, Gary Feng, and Christopher Agard from ETS for support on this work. We also thank the anonymous reviewers of EDM 2020 for their valuable comments on this work.

7. REFERENCES

[1] National Research Council. 1996. *National Science Education Standards*. The National Academies Press, Washington, DC. DOI=https://doi.org/10.17226/4962.

[2] National Assessment Governing Board. 2015. *Science Framework for the 2015 National Assessment of Educational Progress*. Washington, DC. https://www.nagb.gov/naep-frameworks/science/2015-science-framework.html.

[3] National Research Council. 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC. DOI=10.17226/13165.

[4] Bybee, R. W. 2000. Teaching science as inquiry. In *Inquiring into Inquiry Learning and Teaching in Science*, J. Minstrell and E. H. van Zee, Eds. American Association for the Advancement of Science, Washington, D.C., 21–46.

[5] National Research Council. 2013. *Next Generation Science Standards: For States, by States.* The National Academies Press, Washington, D.C. https://www.nap.edu.

[6] Rönnebeck, S., Bernholt, S., and Ropohl, M. 2016. Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies of Science Education*, 52(2), 161–197. DOI=10.1080/03057267.2016.1206351.

[7] Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. 2003. *Design Patterns for Assessing Science Inquiry* (PADI Technical Report 1). Menlo Park, CA. https://padi.sri.com.

[8] Scalise, K. 2014. *Assessment System Design Options for the Next Generation Science Standards (NGSS): Reflections on Some Possible Design Approaches*. ETS, Princeton, NJ. https://www.ets.org/research/policy_research_reports/publications/paper/2014/jvha.

[9] Chen, Z. and Klahr, D. 1999. All other things being equal: acquisition and transfer of the control-of-variables strategy. *Child Development*, 70(5), 1098–1120. DOI=10.1111/1467-8624.00081.

[10] Tschirgi, J. E. 1980. Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51(1), 1–10. DOI= 10.2307/1129583.

[11] Kuhn, D. and Dean, D. 2005. Is developing scientific thinking all about learning to control variables? *Psychological Science*, 16, 866–870. DOI=10.1111/j.1467-9280.2005.01601628.x.

[12] Montgomery, D. C. 2000. *Design and Analysis of Experiments*, 5th ed. Wiley Text Books, Indianapolis, IN.

[13] Black, R. 2007. *Pragmatic Software Testing: Becoming an Effective and Efficient Test Professional*. Wiley, New York.

[14] Klahr, D. and Nigam, M. 2004. The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, 15, 661–667. DOI=10.1111/j.0956-7976.2004.00737.x.

[15] Harrison, A. M. and Schunn, C. D. 2004. The transfer of logically general scientific reasoning skills. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, K. Forbus, D. Gentner, and T. Regier, Eds. Lawrence Erlbaum Associates, Mahwah, NJ, 541–546.

[16] Kim, M. C., Hannafin, M. J., and Bryan, L. A. 2007. Technology-enhanced inquiry tools in science education: An emerging pedagogical framework for classroom practice. *Science Education*, 91(6), 1010–1030. DOI=10.1002/sce.20219.

[17] Shimoda, T. A., White, B. Y., and Frederiksen, J. R. 2002. Student goal orientation in learning inquiry skills with modifiable software advisors. *Science Education*, 86(2), 244–263. DOI=10.1002/sce.10003.

[18] Songer, N. B., Lee, H. S., and Kam, R. 2002. Technology-rich inquiry science in urban classrooms: What are the barriers to inquiry pedagogy? *Journal of Research in Science Teaching*, 39(2), 128–150. DOI=10.1002/tea.10013.

[19] Ebenezer, J., Kaya, O. N., and Ebenezer, D. L. 2011. Engaging students in environmental research projects: Perceptions of fluency with innovative technologies and levels of scientific inquiry abilities. *Journal of Research in Science Teaching*, 48(1), 94–116. DOI=10.1002/tea.20387.

[20] Gobert, J. D., Sao Pedro, M. A., Baker, R. S. J. D., Toto, E., and Montalvo, O. 2012. Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, 4(1), 104–143. DOI=10.5281/zenodo.3554645.

[21] Taasoobshirazi, G., Zuiker, S. J., Anderson, K. T., and Hickey, D. T. 2006. Enhancing inquiry, understanding, and achievement in an astronomy multimedia learning environment. *Journal of Science Education and Technology*, 15(5), 383–395. DOI=10.1007/s10956-006-9028-0.

[22] Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., and Clay-Chambers, J. 2008. Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, 45(8), 922–939. DOI=10.1002/tea.20248.

[23] Minner, D. D., Levy, A. J., and Century, J. 2010. Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, 47(4), 474–496. DOI=10.1002/tea.20347.

[24] Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., and Tsourlidaki, E. 2015. Phases of inquiry-based learning: Definitions and the inquiry cycle. *Education Research Review*, 14, 47–61. DOI=10.1016/j.edurev.2015.02.003.

[25] Wilson, C. D., Taylor, J. A., Kowalski, S. M., and Carlson, J. 2010. The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching*, 47(3), 276–301. DOI=10.1002/tea.20329.

[26] Dostál, J. 2015. Theory of problem solving. *Procedia-Social and Behavioral Sciences*, 174(1), 2798–2805. DOI=10.1016/j.sbspro.2015.01.970.

[27] Dixon, W. J. 1960. Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*, 31(2), 385–391. DOI=10.1214/aoms/1177705900.

[28] Kruskal, W. H. and Wallis, W. A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621. DOI=10.2307/2280779.

[29] R Core Team. 2019. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. http://r-project.org.

[29] Arslan, B., Keehner, M., Jiang, Y., Gong, T., Katz, I. R., and Yan, F. 2020. The effect of drag-and-drop item features on test-taker performance and response strategies. *Educational Measurement: Issues and Practices*. DOI=10.1111/emip.12326.

[30] Shimoda, T. A., White, B. Y., and Frederiksen, J. R. 2002. Student goal orientation in learning inquiry skills with modifiable software advisors. *Science Education*, 86(2): 244–263. DOI=10.1002/sce.10003.