Visualizing depth of student conceptual understanding using subquestions and alluvial diagrams

Jun-ichiro Yasuda*
Institute of Arts and Sciences, Yamagata University, Yamagata, Yamagata 990-8560, Japan

Michael M. Hull
Department of Physics, University of Alaska Fairbanks, 1930 Yukon Dr, Fairbanks, Alaska 99775, USA

Naohiro Mae
Research Center for Nuclear Physics, Osaka University, Ibaraki, Osaka 567-0047, Japan

(Received 20 April 2023; accepted 20 July 2023; published 31 August 2023)

We aim to graphically analyze the depth of conceptual understanding behind the Force Concept Inventory (FCI) responses of students, focusing on three questions (questions 1, 15, and 28). In our study, we created and implemented subquestions to clarify and quantify the students' reasoning steps in reaching their responses to the original FCI questions. We used alluvial diagrams to visualize the responses of students on the subquestions and original FCI questions, representing the transition rates across the question options as flows. The flows represent common patterns of consistency and inconsistency in the students' conceptual understandings across the original FCI question and the subquestions. By combining subquestions and alluvial diagrams as we did, we are able to efficiently analyze the depth of understanding for a given situation and to visualize the existence of a specific misconception. For example, we found that even among those who correctly answered FCI question 15 about Newton's third law, many students incorrectly answered a corresponding subquestion asking about whether Newton's third law holds at the end of the interaction between the two objects. This misconception turned out to be fairly robust, as these same students had a similar answering pattern for question 28. Alluvial diagrams can be used for a broad range of applications, including analysis of multiple-tier tests, testlets (e.g., FCI questions 8 through 11), tests for representational consistency (e.g., a representational variant of the FCI), and so on.

DOI: 10.1103/PhysRevPhysEducRes.19.020121

I. INTRODUCTION

We use concept surveys to identify common misconceptions students hold and design instructional strategies that target these misconceptions. One of the most widely used concept surveys is the Force Concept Inventory (FCI) [1]. The FCI probes students' conceptual understanding of Newtonian mechanics, particularly regarding the concept of force. The test has 30 items with five choices, and the four incorrect answer choices include distractors that are based upon research of students' misconceptions.

In any test, including multiple-choice tests like the FCI, the test taker commits his or her thoughts, whether they be coherent or fragmented, to a symbolic response. It is often the case that the same symbolic response can reflect different depths of understanding. In question 1 (Q.1) of the FCI (discussed in more detail below), for example, students are asked to compare the time necessary for two balls of different masses to reach the ground [1]. Students can answer this question based upon rote memorization that objects hit the ground at the same time. Students can also answer correctly with a deeper conceptual understanding about involving different forces that are perfectly compensated by the masses, resulting in identical accelerations. Many researchers have analyzed what range of reasoning can lie behind student responses on the FCI [1–4]. Qualitative methods (e.g., interviews or open-ended questions) following the tests are commonly used to examine the reasoning behind the responses. Quantitative methods have also been developed to analyze the trends in the reasoning behind the responses. For example, one of the common methods is multiple-tier tests (two-tier, three-tier, and four-tier tests) [5–18]. Depending on the number of tiers, multiple-tier tests may ask students to provide justification for the answer they provided to the original conceptual question, and they may also ask how confident students are in their responses.

Many studies using the above methods focus on false positives, which is when a student answers a question

*Corresponding author: phys.cat.collaboration@gmail.com

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

correctly without understanding the physics concept being tested in the question [1,2]. False positives will appear when a respondent happens to answer correctly by random guessing. For instance, false positives on the FCI appear 20% of the time when a respondent randomly chooses an answer because the FCI is a five-choice survey [2]. False negatives will appear, for example, when a respondent is careless or inattentive; There is much less concern about false negatives, as the probability of their occurrence is relatively small [2].

As an alternative quantitative method to probe student reasoning, we developed subquestions to accompany the original FCI questions [19,20]. (The "sub" in "subquestion" has a similar meaning to its meaning in the word "subatomic" A subquestion is a smaller-sized question.) Each subquestion is designed to test the details of student understanding (see below for examples). Subquestions do not ask students to provide a justification for their selection to the original FCI question; rather, they more thoroughly probe the conceptual reasoning used by the respondent in answering the original FCI question. A set of subquestions can test the logic behind the response to the original FCI question. For example, while an FCI question might ask respondents to describe an object's motion in a given situation, corresponding subquestions can test the student's reasoning by asking about the velocity and acceleration of the object, the force exerted on the object, etc. Such subquestion sets measure the depth of a student's understanding, discerning whether their correct response is coming from superficial memorization or from a more mature understanding. The subquestion sets discussed in this paper are designed to measure the depth of conceptual understanding.

Previous work has used subquestions to analyze false positives and to evaluate the validity of the original FCI questions. We calculated the false positive ratio (false positive occurrence among those who do not understand the physics concept) of the FCI questions and compared them between prelearners and postlearners of Newtonian mechanics, as a means to examine the validity of each FCI question [19]. We also calculated the systematic error (by subtracting the true score from the raw score) due to the false positives and evaluated how much they inflated the total score of the FCI [20]. These analyses were novel as they statistically examined the effect of false positives on the FCI; however, this line of research, which focused on the validity of the questions, did not provide insights into improving pedagogy. This is a gap in previous literature that this paper aims to fill. Here, we will discuss how subquestions can be used for informing the development and improvement of curriculum.

In this paper, we suggest another use of subquestions that is not for subsequent statistical computation but rather for the graphical visualization of students' conceptual understanding. The method of visualization is one of the methodologies in physics education research, supported by Wittmann *et al.* [21] who stated that "Good representations can provide answers to difficult questions and raise questions that would not have been previously asked." Wittmann *et al.* introduced the representation of consistency plots, which is an extension of Kanim's escalator diagrams showing changes in student responses from correct to incorrect while representing preinstruction and postinstruction. In another study on visualization, Tapping *et al.* [22] generated networklike graphs for each question of Conceptual Survey of Electricity and Magnetism, where responses at pretest and post-test are represented by nodes connected by edges representing the change in student response choice. Note that these studies compared pretest and post-test responses, but did not compare responses of students on the individual level (for example, within the pretest) with each other to look for depth of understanding as we do in this paper.

In our analysis, we use alluvial diagrams [23] to visualize trends in students' reasoning steps behind the FCI responses. (An alluvial diagram is a subcategory of a Sankey diagram; the details of the difference is explained in [24]). Specifically, in using alluvial diagrams with subquestions, we represent consistency and inconsistency in students' conceptual reasoning across the original FCI question and their subquestions. Alluvial diagrams are graphs to simplify and highlight the structural changes between clusters so that any transition rate can be seen at a glance. Compared with analysis using tables of transition rates, alluvial diagrams improve the efficiency of the analysis, improve the readability of the representation, and prevent oversight. Alluvial diagrams are used in a variety of research fields, for example, in physics education research, Traxler *et al.* [25] used alluvial diagrams to show to what extent students stay in the same social group from week 1 to week 10 of the term. Werth *et al.* [26] used an alluvial diagram to show the change in learning goals of instructors from before to after remote instruction.

In item analysis, alluvial diagrams are used to visualize transition rates between item responses. For example, Atkinson and Bretz [27] used alluvial diagrams to show the percentage of students choosing each answer to the Reaction Coordinate Diagram Inventory on the pretest and post-test, focusing on the changes in choosing attractive distractors. Doherty *et al.* [28] used alluvial diagrams in a similar manner in neurophysiological education research. There are other studies that use alluvial diagrams to analyze the transition between the pre and post-test, but we found no studies, despite a thorough search, that use alluvial diagrams to analyze transitions in responses to a series of questions.

In this study, we created and implemented new subquestions to the FCI that enable us to glean insight into how students were reasoning about the original corresponding FCI items. We analyze the trends in student reasoning about the FCI questions and associated subquestions using alluvial diagrams. We address the following research questions:

1. How can we interpret trends in student reasoning behind the FCI questions using subqestions and alluvial diagrams?
2. What are the implications for improving teaching and assessment methods suggested by the analysis?

The remainder of this paper is organized as follows. In Sec. II, we describe the methodology of this study, the subquestions we used in this study, and how we created those subquestions. The design of the survey in this study is also described. In Sec. III, we present the results of our analysis using alluvial diagrams. In Sec. IV, we describe the implications elicited from the results. Finally, in Sec. V, we summarize this study and discuss the prospects for it.

II. METHODOLOGY

The objective of this paper is to analyze the conceptual understanding behind the FCI responses of students. In the analysis, we use subquestions to clarify and quantify the student's reasoning steps to reach the FCI response. We calculate and compare the percentage of students who selected each reasoning step in an item unit (recall that an "item unit" is an FCI question with its subquestions). Furthermore, we use alluvial diagrams to visualize the percentage of students who chose each possible reasoning step and, more importantly, the transition rate between the item responses in a series of questions. This diagram represents the percentage of students selecting each response for a given item as a bar plot and the transition rate between the responses as a flow (see below for examples). The flows represent common patterns of consistency and inconsistency in the students' conceptual understandings across the original FCI question and the subquestion(s). Combining subquestions and alluvial diagrams, we can more deeply analyze the students' reasoning behind the FCI response and can visualize the existence of specific misconceptions. To produce alluvial diagrams, we used the R package ggalluvial [24], which is an extension for ggplot2.

A. Subquestions

The design of the subquestions is primarily based on student think-aloud interviews in our previous study (see [29] for details on the interview protocol and results). If we observed a false positive or a false negative with incorrect reasoning for a given item, we created a series of questions, where each question tests the details of student understanding. When neither false positive nor false negative were found for a given item, one can consider what the original FCI item was intended to assess, by consulting the FCI taxonomy chart [30], the misconceptions table [31], or the true reasoning interviewees stated [32]. As such, we utilized a variety of resources to develop subquestions for the FCI questions.

In our previous work, we documented how we created and analyzed results from subquestions of four FCI items [19,20]. In 2022, we newly developed the subquestions of the rest of the FCI questions. Among these new subquestions, in this paper, we chose to focus on Q.1, Q.15, and Q.28 because we found implications to improve the teaching or assessment method from the analysis of alluvial diagrams with these questions. In the following section, we describe how we developed the subquestions for these FCI questions.

1. Question 1

The first question we focused on is Q.1 of the FCI. The description of this question is shown in Item unit 1.

Item unit 1.

Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the roof of a single story building at the same instant of time.

Q.1 The time it takes the balls to reach the ground below will be:

(1) about half as long for the heavier ball as for the lighter one.
(2) about half as long for the lighter ball as for the heavier one.
(3) about the same for both balls.
(4) considerably less for the heavier ball, but not necessarily half as long.
(5) considerably less for the lighter ball, but not necessarily half as long.

SQ.1.1 How do the accelerations of the two balls compare?

(1) The heavy ball's acceleration is about half that of the light ball.
(2) The light ball's acceleration is about half that of the of heavy ball.
(3) They are about the same.
(4) The heavy ball's acceleration is quite a bit smaller, but not quite half as small as the light ball.
(5) The light ball's acceleration is quite a bit smaller, but not quite half as small as the heavy ball.

SQ.1.2 The Earth pulls on the light ball ⋯

(1) ⋯ half as hard as on the heavy ball.
(2) ⋯ twice as hard as on the heavy ball.
(3) ⋯ the same as it does on the heavy ball.
(4) ⋯ less hard than it does on the heavy ball, but not quite half as much.
(5) ⋯ harder than it does on the heavy ball, but not quite twice as hard.

This question (Q.1) tests "acceleration independent of weight" for "Gravitation" [30]. For this question, one interviewee (S1) in our previous study [29] selected the right answer with the following reasoning in the interview with one of the authors (I):

S1: *"I chose (3) because I thought of something like that thought experiment, or dropping it off the Leaning Tower of Pisa."*

I: *"What was the Leaning Tower of Pisa experiment like?"*

S1: *"In the end, they dropped metal balls of different weights together, but they arrived at the same time."*

This student just recalled that "balls of different weights reach the ground at the same time" from the anecdote of Galileo's experiment, without giving any evidence of understanding the relationship between gravitational acceleration and weight. Inspired by this student's answer, we designed the subquestions of Q.1 shown in Item unit 1 (SQ.1.1 and SQ.1.2). These subquestions ask about the magnitude of the accelerations of and the gravitational forces on the two balls.

2. Question 15

The second question we focused on is Q.15 of the FCI. The description of this question is shown in Item unit 15.

This question (Q.15) tests "Newton's third law for continuous forces" [30]. For this question, one interviewee (S2) in our previous study [29] selected the right answer with the following reasoning:

S2: *"Umm, the two cars are moving at the same speed while being completely in contact, so if one of them is not in balance with the other, umm, huh? Yes, they're moving together, exactly together, so I thought that the two forces, the pushing force and the pushing back force, are the same."*

Since this student explained that the magnitude of the action and reaction forces are the same because "the car and truck are in contact," we thought that this student might believe that the magnitude of the action and reaction forces would not be the same depending on the state of motion of the car and truck. Therefore, we designed a subquestion for Q.15 to ask about the magnitudes of the forces exerted by the truck on the car and by the car on the truck at the moment of separation (see SQ.15.1 in Item unit 15).

Item unit 15.

A large truck breaks down out on the road and receives a push back into town by a small compact car as shown in the Fig. 1.

Q.15 While the car, still pushing the truck, is speeding up to get up to cruising speed:

(1) The amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car.
(2) The amount of force with which the car pushes on the truck is smaller than that with which the truck pushes back on the car.
(3) The amount of force with which the car pushes on the truck is greater than that with which the truck pushes back on the car.
(4) The car's engine is running so the car pushes against the truck, but the truck's engine is not running so the truck cannot push back against the car. The truck is pushed forward simply because it is in the way of the car.
(5) Neither the car nor the truck exert any force on the other. The truck is pushed forward simply because it is in the way of the car.

FIG. 1. The figure of the FCI Q.15.

SQ.15.1 The car that is continuously pushing the truck decelerates at a certain instant and pulls away from the truck. In this case, which of the following is true about the forces exerted by the car on the truck and by the truck on the car between the time the car and the truck are pushing and the moment the car leaves?

(1) The car always exerts a greater force.
(2) The truck always exerts a greater force.
(3) During pushing, the forces exerted are of the same magnitude, but at the moment of separation, the force exerted by the car is greater than that exerted by the truck.
(4) During pushing, the forces exerted are of the same magnitude, but at the moment of separation, the force exerted by the truck is greater than that exerted by the car.
(5) The car and truck always exert the same magnitude of force on each other.

3. Question 28

In order to confirm the misconception about Newton's third law not being applicable in a moment of separation, we also focused on Q.28 and created a subquestion in a similar manner as Q.15. The original Q.28 is in Item unit 28.

This question (Q.28) tests "Newton's third law for impulsive forces" [30]. For this question, one interviewee (S3) in our previous study [29] selected the right answer with the following reasoning:

S3: *"In this case, since student "a" is touching student "b," I thought that one of them would move away from each other immediately if they did not exert the same magnitude of force, so I thought they exerted the same magnitude of force on each other."*

I: *"So if the magnitude of the force is different, they would move away from each other?"*

S3: *"Yes, even if the magnitude of the force is the same, huh? But, … I didn't really understand this question either. Well, I thought that if "a" exerts a force on "b" in an action-reaction relationship, then "b" exerts a force of the same magnitude on "a" so I chose (5)."*

I: *"Your explanation just now was different from the first one you said, though, wasn't it? Did you just now remember the action-reaction?"*

S3: *"Yes."*

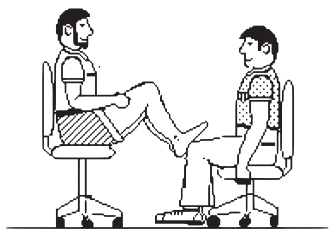

FIG. 2. The figure of the FCI Q.28.

At first, the interviewee explained that the magnitude of the action and reaction forces were the same because different magnitudes would separate the student "a" and "b." Later, this interviewee reconsidered and based the explanation on Newton's third law. We considered this interviewee's initial statement as being similar to that of the interviewee discussed above for Q.15 and thought that this interviewee might also believe that the magnitude of the action and reaction forces might not be the same under certain circumstances. Therefore, the subquestion of Q.28, as in Q.15, was designed to ask for the magnitudes of the forces at the moment when students "a" and "b" leave each other (see SQ.28.1 in Item unit 28).

Item unit 28.

In the Fig. 2, student "a" has a mass of 95 kg and student "b" has a mass of 77 kg. They sit in identical office chairs facing each other. Student "a" places his bare feet on the knees of student "b," as shown. Student "a" then suddenly pushes outward with his feet, causing both chairs to move.

Q.28 During the push and while the students are still touching each other:

(1) Neither student exerts a force on the other.
(2) Student "a" exerts a force on student "b," but "b" does not exert any force on "a."
(3) Each student exerts a force on the other, but "b" exerts the larger force.
(4) Each student exerts a force on the other, but "a" exerts the larger force.
(5) Each student exerts the same amount of force on the other.

SQ.28.1 From the moment when the foot is still on the knee during the push to the moment when the foot leaves the knee,

(1) student "a" always exerts a larger force.
(2) student "b" always exerts a larger force.
(3) during pushing, both exert the same magnitude of force, but at the moment of separation, student "a" exerts a larger force.
(4) during pushing, both exert the same magnitude of force, but at the moment of separation, student "b" exerts a larger force.
(5) they always exert the same amount of force on the other.

B. Survey design

The questions in each item unit (the original FCI question and its series of subquestions) were arranged in a sequence that retroactively asks for logic so as to minimize hints given by the previous questions. The test was administered by computer and students were prevented from going back to the previous question to change their answer. This is an improvement from our previous study [19,20], where we surveyed with scantrons and simply instructed students not to return to the previous questions.

Among the subquestions we created in this study, we chose to survey the item units of FCI Q.1, Q.2, Q.8, Q.12, Q.14, Q.15, Q.23, Q.27, and Q.28. These are the items for which we found false positives or false negatives in our previous interview study but had not yet made subquestions in our previous work. We validated the subquestions of these items through a pilot survey. We analyzed the survey data and did not find any anomalous results. Furthermore, we conducted think-aloud interviews with two students (one with a 97% correct response rate and the other with a 65%) to ask their reasoning behind answering each question. We acknowledge that one round of think-aloud interviews with only two students is less than ideal. Our claims that the subquestions are valid would have been strengthened had we interviewed more students; however, since these subquestions are derived from the well-validated FCI, we expected that the survey validation process would not need to be as extensive as when one creates new survey questions from scratch. Based on the interview results, we removed or modified some questions. For example, we removed a subquestion for FCI Q.12, for which an interviewee answered with testwiseness [4] saying "Since this question is the same as the previous question only with an additional condition, I assumed the answer would be different from that question." We also removed a subquestion for FCI Q.14 which was so similar to the original FCI Q.14 that both interviewees reported that they did not find a difference. In addition to these changes, we modified the sentences of subquestions where students indicated that the meaning of the questions was not clear. We did not make any modifications to the subquestions we described above (Q.1, Q.15, and Q.28), which are the FCI items we focus upon in this paper.

The final version of the survey sheet is outlined in Table I. The number of item units was 9 and the number of subquestions was 18 in total after removing two subquestions as we described above. We included an additional 3 questions from the FCI without subquestions, giving us a total of 12 items with which to equate [33] the surveyed items to the 30 original FCI items. (Equating in

TABLE I. Design of the survey sheet. The original FCI Q.13, 20, and 30 are for test equating.

Item unit	Original FCI	No. of SQs
1	1	2
2	2	3
8	8	3
12	12	2
(for equating)	13	0
14	14	2
15	15	1
(for equating)	20	0
23	23	1
27	27	2
28	28	2
(for equating)	30	0

item response theory is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably as in [33].) In total, the survey questions consisted of 12 original FCI items and 18 subquestions.

The main survey was conducted in 2022 in physics classes (including lectures and labs) at five Japanese universities. Of the five universities, four were mid-level and one was a top-level university. Because the class of the top-level university in which the survey was administered was for nonphysics majors, the student's proficiency level in mechanics was, on average, comparable to the average proficiency levels in the other classes.

The survey was conducted online. The total number of respondents was 457. We cleaned the survey data based on the method of visual inspection of response time distributions [34] as in the following steps. First, we compiled the response times for each question into a single dataset for all questions and all respondents and drew a histogram. Second, we examined the dip of the bimodal distribution, which we interpret as the point at which the rapid guessing and solution behavior distributions intersect in the response time distribution. This dip occurred at 5 s, and so we assume that response times less than 5 s correspond to rapid-guessing behavior. We excluded all responses of the respondents who exhibited any rapid-guessing behavior. As a result, 359 responses were obtained as valid responses. As a more precise but complicated method, we also tried to calculate the response time distribution of each question based upon the dip item by item. However, this did not affect our conclusions significantly.

III. RESULT

Figure 3 (left) shows the alluvial diagram for the item unit of the FCI Q.1. The stacked bar plot in the graph shows the percentage of students selecting each option for each item (Q.1, SQ.1.1, and SQ.1.2), and the width of the flow shows the percentage of students selecting combinations of options in Q.1, SQ.1.1, and SQ.1.2. (We refer to the width of the flow as the "transition rate"). The percentages of selecting each option or flow, the values shown in the vertical axis, are calculated among the total valid responses ($N_v = 359$).

In order to interpret the results, we hide flows of ratios less than 3% in Fig. 3 (right). In this graph, the flow of completely correct answers is (3)-(3)-(1): 34.0%, which is selected by the largest percentage of students. The next largest percentage flow is (3)-(3)-(3): 21.4%. The students who selected this flow think that the accelerations of the two balls with different weights are the same, and the gravitational force exerted is also the same. These students

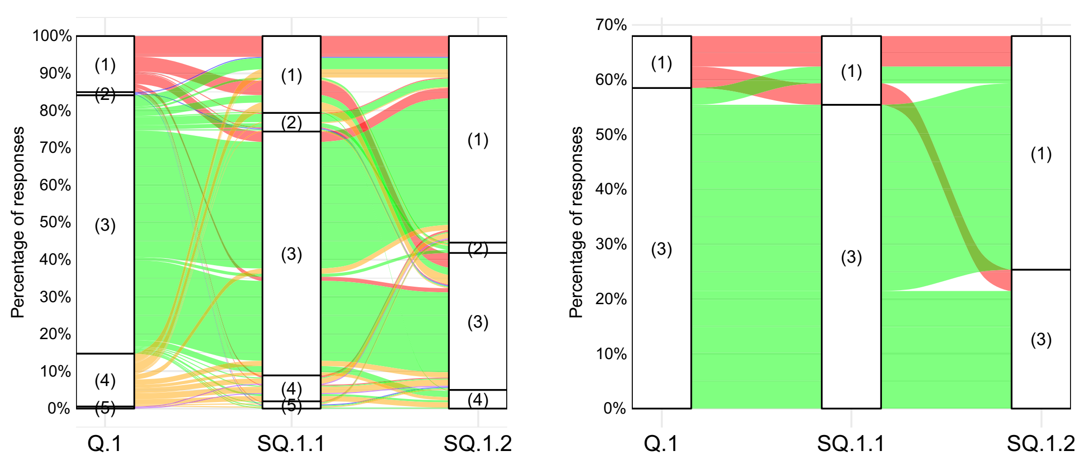

FIG. 3. (Left) Alluvial diagram for the item unit of FCI Q.1. The selection rate for each option is shown as a bar plot, and the transition rate between responses is shown as the width of the flow. The percentages of selecting each option or flow, the values shown in the vertical axis, are calculated among the total valid responses ($N_v = 359$). The path correctly answered in the leftmost question is colored in green, and the paths incorrectly answered in the leftmost question are in red, orange, blue, and purple in descending order of selection rate. (Right) Alluvial diagram for the item unit of FCI Q.1 with flows with ratios less than 3% hidden.

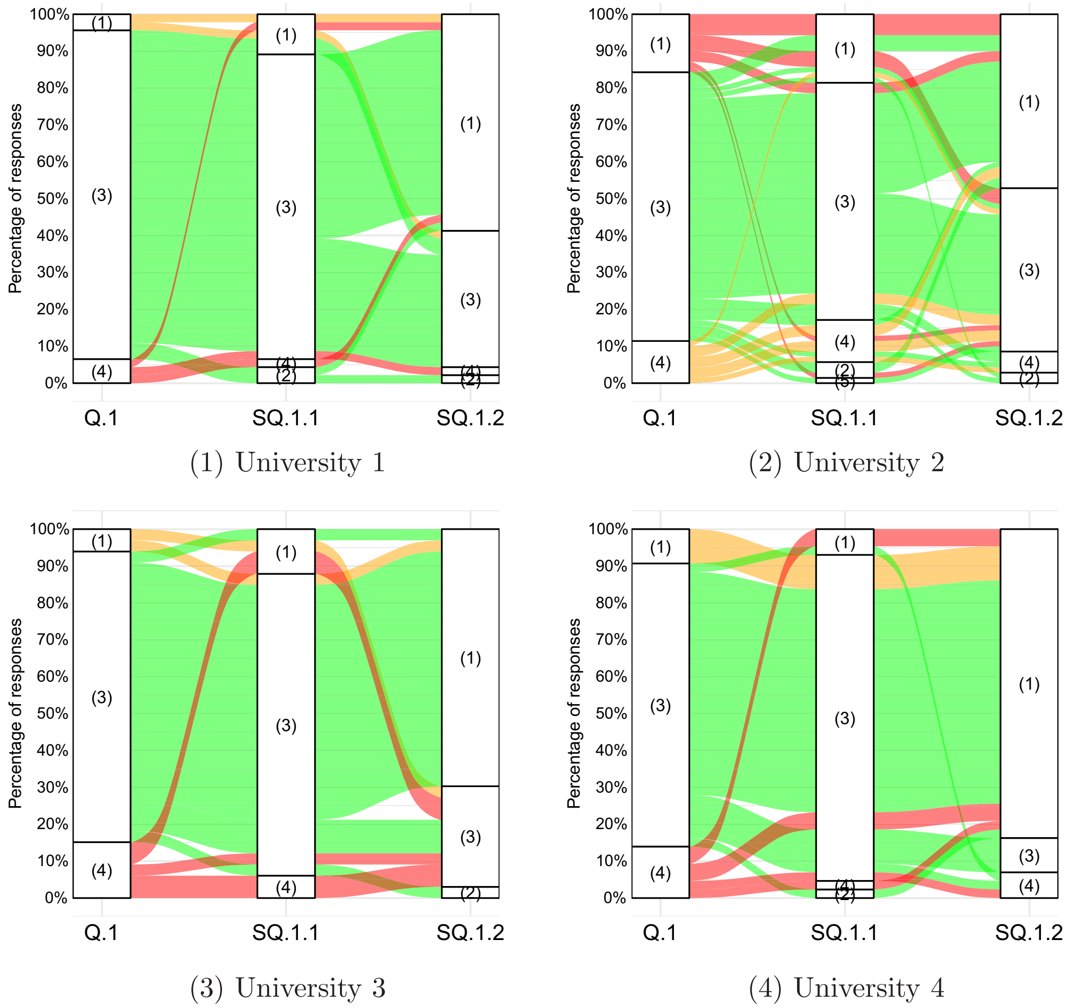

FIG. 4. Alluvial diagrams for the item unit of FCI Q.1 for different universities. The percentages of selecting each option or flow, the values shown in the vertical axis, are calculated among the number of valid responses for each University (1) $N_v = 46$, (2) $N_v = 70$, (3) $N_v = 33$, and (4) $N_v = 43$, respectively. The color scheme is the same as in Fig. 3. The options in the bar chart are (automatically with R) ordered to make it easier to see the flow boundaries.

perhaps have the misconception of "conflating acceleration and force." The next largest percentage flow is (1)-(1)-(1): 5.6%. The students who selected this flow would have the misconception of "Heavier objects fall faster" in the taxonomy table [31]. These students also conflate acceleration and force, thinking that the greater gravitational force on the heavier ball implies a greater acceleration and thus a shorter fall time. For the other flows in Fig. 3 (right), (1)-(1)-(3): 3.9% or (3)-(1)-(1): 3.1%, we cannot find a reasonable explanation for their reasoning with Newtonian conceptions and/or misconceptions.

Overall, it seems that the broad flows in the alluvial diagram correspond to Newtonian or misconception-based reasoning and the narrow flows indicate inconsistent, fragmented reasoning. This may, however, be due in part to the large percentage of students who answered the original FCI question correctly. With lower proficiency groups, the Newtonian-based reasoning flow will be narrower and may be of comparable width to the fragmented reasoning flows. However, with middle and high proficiency groups, the width of alluvial diagram flows may be a means of distinguishing self-consistent reasoning from reasoning that is more fragmented. Based on this inference, we compare the flows from Q.1 (1) and (4) in Fig. 3 (left). Although the selection rates of Q.1 (1) and (4) are almost the same, there are many narrow flows out of (4). This may indicate that the students who selected (4) tend to have a fragmented understanding on the physics concept tested in Q.1.

In Fig. 4, from the five universities we surveyed, we chose four universities that have similar class sizes from which to compare alluvial diagrams. We can compare the depth of understanding of these four student groups. For Q.1 of the original FCI, the percentage of students answering correctly is high in all four groups. This means that most students at these universities at least *know* that

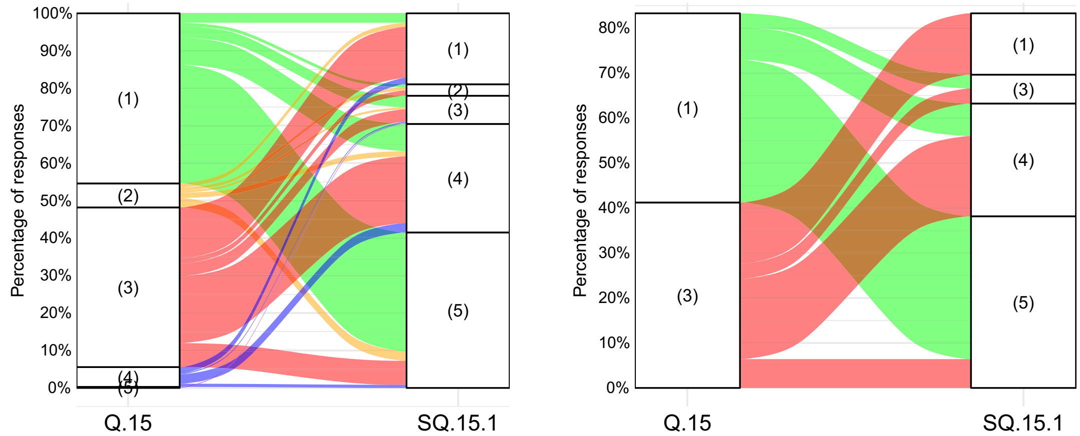

FIG. 5. (Left) Alluvial diagram for the item unit of FCI Q.15. The percentages shown in the vertical axis are calculated among the total valid responses ($N_v = 359$). The color scheme is the same as in Fig. 3. (Right) Alluvial diagram for the item unit of FCI Q.15 with flows with ratios less than 3% hidden.

two objects of different masses fall in the same time. For SQ.1.1, the percentage of students answering correctly is also quite high in all groups. This means that most students at these universities at least *know* that the gravitational acceleration of two objects with different masses is the same. For SQ.1.2, however, the percentage of students answering correctly at University 3 and at University 4 is notably higher than that at University 1 and at University 2. This indicates that the students at University 3 and at University 4 understand that the heavier object experiences a greater gravitational force on it than the students at University 1 and at University 2.

In this way, by using subquestions and alluvial diagrams, we can analyze and compare the depth of understanding of a student in a given situation (in this case, two falling balls). Alluvial diagrams, like tables of transition rates, allow researchers to find patterns in their data. Alluvial diagrams have the benefit of allowing analysis by reading overall trends without having to read the numbers.

Figure 5 (left) shows the alluvial diagram for the item unit of the FCI Q.15. In Fig. 5 (right), flows with a ratio of less than 3% were hidden. In the graph, the flow of completely correct answers is (1)-(5): 31.8%, which is selected by the largest percentage of students. The second and third largest percentage flows are (3)-(4): 17.8%, and (3)-(1): 13.6%. The students who selected these flows would have the misconception of "Most active agent produces greatest force" [31], although, at the moment of separation, there seems to be a split in judgment as to which (car or truck) is the most active agent. Albeit less prevalent, two noteworthy flows are (1)-(3): 3.1%, and (1)-(4): 7.2%. These students answered the original FCI question correctly; however, they incorrectly answered the subquestion. The total percentage of these two flows is

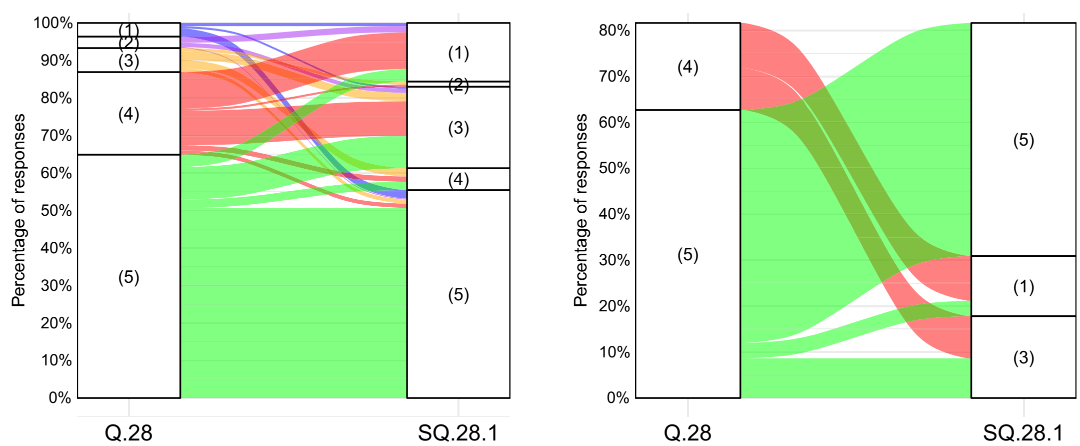

FIG. 6. (Left) Alluvial diagram for the item unit of FCI Q.28. The percentages shown in the vertical axis are calculated among the total valid responses ($N_v = 359$). The color scheme is the same as in Fig. 3. (Right) Alluvial diagram for the item unit of FCI Q.28 with flows with ratios less than 3% hidden. The options in the bar chart are (automatically with R) ordered to make it easier to see the flow boundaries.

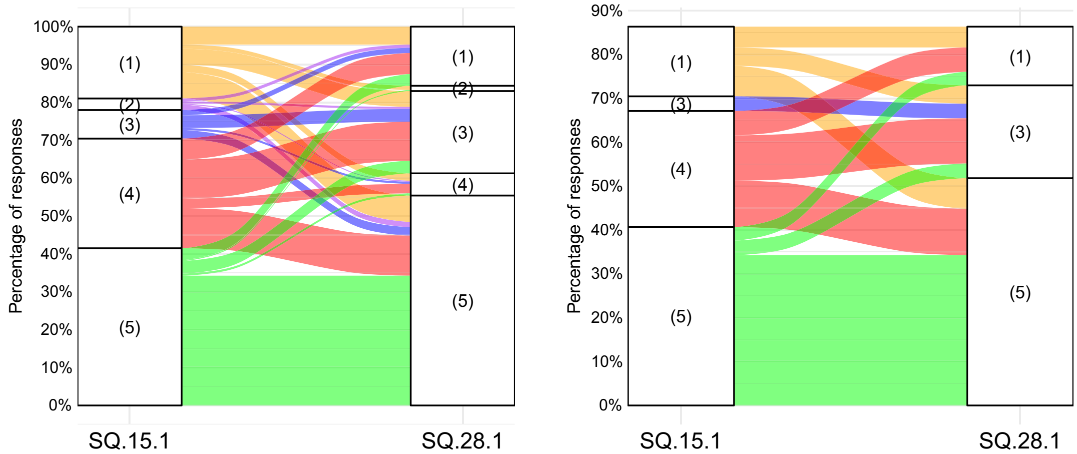

FIG. 7. (Left) Alluvial diagram for the subquestions of Q.15 and Q.28. The percentages shown in the vertical axis are calculated among the total valid responses ($N_v = 359$). The color scheme is the same as in Fig. 3. (Right) Alluvial diagram for the subquestions of Q.15 and Q.28 with flows with ratios less than 3% hidden.

10.3% among valid respondents and 22.7% among respondents who correctly answered Q.15 [selected option (1)]. These students answered Q.15 correctly, but they believe that one force is greater than the other at the moment of separation. It seems that these students believe that Newton's third law does not hold in certain circumstances. We will discuss this misconception in a later section. The flow (3)-(5): 6.4% is difficult to interpret. Here, although students answered the original FCI item incorrectly, they answered the subquestion correctly. The reasoning of these students might be particularly fragmented.

Figure 6 (left) shows the alluvial diagram for the item unit of the FCI Q.28. In Fig. 6 (right), flows with a ratio of less than 3% are hidden. In the graph, the flow of completely correct answers is (5)-(5): 50.7%, which is selected by the largest percentage of students. The second and third largest percentage flows are (4)-(1): 9.7% and (4)-(3): 9.2%. The students who selected these flows are answering the original FCI question with the response consistent with the misconception of "Greater mass implies greater force" and/or "Most active agent produces greatest force" [31]. The students in the second flow, however, do not answer the subquestion in a way consistent with this misconception, demonstrating unstable reasoning. A notable flow is (5)-(3): 8.6%. Although these students answered the original FCI question correctly, they incorrectly answered the subquestion. These students answered Q.28 correctly, but they believe that the force by student "a" is greater at the moment of separation. These students seem to believe that Newton's third law does not hold in certain circumstances. The students with the flow of (5)-(1): 3.3% also answered Q.28 correctly, but their reasoning is unstable, as they answered the subquestion in a way that directly contradicts their answer to the original FCI question.

Figure 7 (left) shows the alluvial diagram for the subquestions of Q.15 and Q.28. In Fig. 7 (right), flows with a less than 3% ratio are hidden. In the graph, the flow of completely correct answers is (5)-(5): 34.3%, which is selected by the largest percentage of students. The second largest percentage flow is (4)-(5): 10.6%. Students who chose this flow answered incorrectly in situations where both objects were moving (Q.15), but correctly in situations where both objects were at rest (Q.28). The third largest percentage flows is (4)-(3): 10.3%. The students who selected this flow consistently believe that Newton's third law does not hold just before two objects separate. In particular, we can see that these students would activate in their minds the misconception of "Greater mass implies greater force" (regardless of whether the object with greater mass was the initiator of the motion) at the end of the interaction.

IV. DISCUSSION

Above, in analyzing the responses of individual students to SQ.15.1 and SQ.28.1, we saw that more than 20% of students selected options, indicating the belief that Newton's third law does not hold just before two objects separate. Specifically, the sum of the percentage of students selecting (3) or (4) on SQ.15.1 is 36.5%, and the percentage on SQ.28.1 is 27.6%. It is particularly noteworthy that many of the students who correctly answered FCI Q.15 and Q.28 responded this way on the subquestions. For example, 10.3% of students answered Q.15 correctly but, on the corresponding subquestion, seemed to believe that Newton's third law does not hold just before two objects separate. Moreover, 16.7% [the sum of the percentage who selected the flows of (3)-(3), (3)-(4), (4)-(3), and (4)-(4) in SQ.15.1 and SQ.28.1] of students consistently answered that Newton's third law does not hold at the end of the interaction of the two objects. These results imply the robustness of this incorrect idea. As this idea appears to be at least somewhat stable and widespread, we would say that this reasoning is a misconception referring to the definition of [35] that a misconception is "a particular mental model or line of reasoning [that] is robust and found in a significant fraction (say on the order of 20% of students or more)."

It is well known that novice learners may think that Newton's third law only holds under certain circumstances. For example, novice learners may think that Newton's third law holds when two objects are at rest, but not when both objects are moving. We have found that student understanding about Newton's third law is yet more nuanced: Even students who understand that Newton's third law applies to two moving objects of different masses (answered Q.15 correctly) are still prone to think that Newton's third law does not apply at the end of an interaction. Despite a thorough review of the literature ([36–39] and references therein), it seems that this misconception has not been documented prior to our work.

Why do students have such a misconception? At this point, all we can do is speculate, informed by the data we have. In the survey validation interview, interviewee S2 answered Q.15 by first (inappropriately) bringing up the idea of forces balancing in saying "*the two cars are moving at the same speed while being completely in contact, so if one of them is not in balance with the other* $\cdots$". Here, S2 seemed to recognize a flaw in the reasoning and abruptly changed track and provided a new justification for the forces being equal: *"umm, huh? Yes, they're moving together, exactly together, so I thought that the two forces, the pushing force and the pushing back force, are the same."* (For Q.28, S3 also used similar reasoning in saying *"I thought that one of them would move away from each other immediately if they did not exert the same magnitude of force, so I thought they exerted the same magnitude of force on each other."*) This is not altogether surprising, as it is known that novice learners misapply Newton's second law instead of applying Newton's third law when two objects interact [40,41], and add the two forces that form the Newton's third law pair to obtain a net force [42]. Previous research has argued that the idea of balancing forces tends to be more salient to learners than Newton's third law [41]. It is plausible that novice learners, drawing upon the notion of force balances, suppose that the reason for objects to separate after a collision is an unbalance of forces, and that the strength of this reasoning overwhelms their recollection of Newton's third law [41]. Again, this is only speculation, and other possible mechanisms may better explain this misconception. An alternative explanation is that students think that the forces become unbalanced when objects separate because they are reasoning using the velocity-dependence facet [37]. When the two objects travel together, they have the same velocity, and so the forces exerted are the same. Students reason that when the objects separate, the velocities become different, and so too must the forces.

How can we remedy this misconception? As one of the five ideas important in careful consideration of the third law, Brown and Clement [43] stated that "At all moments of time the force A exerts on B is of exactly the same magnitude as the force B exerts on A." In order to demonstrate that statement, one well-known method is to use two carts fitted with force sensors to graphically show the magnitude of the force with which the carts push against each other [44]. However, if it is indeed the case that students assume the forces to not be equal when objects are separating because they are utilizing cognitive resources about force balances, we suspect that this approach would be insufficient. In order that the understanding of Newton's third law is not overwhelmed by the misapplication of Newton's second law, it is important to teach the difference between force balancing and action or reaction forces as described in [41].

V. SUMMARY

We developed a method utilizing subquestions and alluvial diagrams to visualize and more deeply analyze student reasoning. Based on this method, we examined the reasoning behind the FCI responses of students, focusing on Q.1, Q.15, and Q.28. One of the advantages of using subquestions and alluvial diagrams is that the depth of understanding can be efficiently analyzed in a given situation and another advantage is ease in visualizing the existence of a specific misconception. The use of alluvial diagrams improves the efficiency of the analysis and the readability of the representation. Our methodology yielded insights into improving pedagogy; namely, we found the misconception that Newton's third law holds only in the middle of an interaction, not when the interaction is ending. Teachers can utilize alluvial diagrams in the way we demonstrated to effectively probe inconsistencies in the conceptual reasoning of their students. Based upon how stable student reasoning is, teachers can consider what style of instruction would be ideal [45]. The respondents of our survey are exclusively Japanese students, so future work should investigate the presence of this misconception in other countries as well. The range of potential applications for alluvial diagrams in analyzing item series is broad; for example, researchers can use it to analyze multiple-tier tests, testlet (e.g., FCI questions 8 through 11), tests for representational consistency [46] and so on. Alluvial diagrams would make it possible to "provide answers to difficult questions and raise questions that would not have been previously asked" [21] in these testing methods.

Our tips for using alluvial diagrams in item analysis are as follows. First, we suggest keeping the number of questions to analyze 4 or fewer items, because if the number of questions in a diagram is larger, the number of flows will increase, and the width of the flows will become narrower. This complicates the interpretation of the results. We recommend focusing on the key questions with

4 or fewer items from which meaningful implications can be elicited. Second, we recommend hiding the narrow flows when interpreting the flows. Although we consistently used the threshold of 3%, it will depend on the number of items analyzed. For analyses with more than 3 items, it may be better to use a smaller threshold value, since the number of flows will be higher, and the flow will be narrower for that number of items. Third, we found a trend that the broad flows in an alluvial diagram can be related to Newtonian and/or misconception reasoning, and the narrow flows to fragmented reasoning. This could be a clue to interpreting alluvial diagrams except when analyzing low-proficiency groups, as discussed above. The researcher should examine the trend of data by themselves considering the distribution of the proficiency level of their examinees. Fourth, for many researchers, analysis using alluvial diagrams is easier compared with more computational statistical methods. Although an online application [47] to draw alluvial diagrams is available, researchers are recommended to draw alluvial diagrams using R. In order to support the researchers who are not familiar with R, we share a sample code to draw an alluvial diagram on our GitHub page [48].

By using a visualization approach, we aimed to take an intermediate position between the qualitative approach and the quantitative approach. Our approach aimed to generalize the finding in our qualitative approach (interviews) but does not aim to be based on methods as formal as quantitative approaches, such as statistical estimation and testing. Even with quantitative methods, sample-dependent issues inevitably remain. We took our analysis on our limited (but not small) sample as an intermediate-scale study to see what we could interpret from it. More general analysis can be done with more samples with formal statistical analysis.

Our inspiration for creating subquestions came from false positives and false negatives detected in student interviews. This approach has the disadvantage of being sample-dependent; namely, it is possible that a different sample of students would have answered with a false positive or false negative that we did not detect in our study. One major limitation of using alluvial diagrams is that, as we described above, it is difficult to interpret the results for many sequential questions (say, more than four questions).

The subquestions we created for this study can be used for another study. Recently, the use of computerized adaptive testing (CAT) is becoming more popular for conceptual surveys [49,50]. Another possibility is using multistage testing instead of CAT, which evaluates a respondent's proficiency in each stage comprising a set of questions. The subquestions together with the corresponding original FCI question can form such a question set, and we intend to use our subquestions to develop such a testing method in future work.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grants No. JP19K03135 and No. JP22H01061.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[2] D. Hestenes and I. Halloun, Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[3] N. S. Rebello and D. A. Zollman, The effect of distracters on student performance on the Force Concept Inventory, Am. J. Phys. **72**, 116 (2004).

[4] S. DeVore, J. Stewart, and G. Stewart, Examining the effects of testwiseness in conceptual physics evaluations, Phys. Rev. Phys. Educ. Res. **12**, 020138 (2016).

[5] D. K. Gurel, A. Eryilmaz, and L. C. McDermott, A review and comparison of diagnostic instruments to identify students' misconceptions in science, Eurasia J. Math. Sci. Technol. Educ. **11**, 989 (2015).

[6] I. Testa, S. Galano, S. Leccia, and E. Puddu, Development and validation of a learning progression for change of seasons, solar and lunar eclipses, and moon phases, Phys. Rev. ST Phys. Educ. Res. **11**, 020102 (2015).

[7] P. Barniol and G. Zavala, Mechanical waves conceptual survey: Its modification and conversion to a standard multiple-choice test, Phys. Rev. Phys. Educ. Res. **12**, 010107 (2016).

[8] P. Klein, A. Muller, and J. Kuhn, Assessment of representational competence in kinematics, Phys. Rev. Phys. Educ. Res. **13**, 010132 (2017).

[9] Y. Xiao, J. Han, K. Koenig, J. Xiong, and L. Bao, Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning, Phys. Rev. Phys. Educ. Res. **14**, 020104 (2018).

[10] L. Bao, Y. Xiao, K. Koenig, and J. Han, Validity evaluation of the Lawson classroom test of scientific reasoning, Phys. Rev. Phys. Educ. Res. **14**, 020106 (2018).

[11] U. ScottidiUccio, A. Colantonio, S. Galano, I. Marzoli, F. Trani, and I. Testa, Design and validation of a two-tier questionnaire on basic aspects in quantum mechanics, Phys. Rev. Phys. Educ. Res. **15**, 010137 (2019).

[12] J.-P. Burde and T. Wilhelm, Teaching electric circuits with a focus on potential differences, Phys. Rev. Phys. Educ. Res. **16**, 020153 (2020).

[13] H.-S. Wang, S. Chen, and M.-H. Yen, Effects of metacognitive scaffolding on students' performance and confidence judgments in simulation-based inquiry, Phys. Rev. Phys. Educ. Res. **17**, 020108 (2021).

[14] L. Ivanjek, L. Morris, T. Schubatzky, M. Hopf, J.-P. Burde, C. Haagen-Schutzenhofer, L. Dopatka, V. Spatz, and T. Wilhelm, Development of a two-tier instrument on simple electric circuits, Phys. Rev. Phys. Educ. Res. **17**, 020123 (2021).

[15] H. Şahin Kızılcık, M. Aygün, E. Şahin, N. Önder Çelikkanlı, O. Türk, T. Taşkın, and B. Güneş, Possible misconceptions about solid friction, Phys. Rev. Phys. Educ. Res. **17**, 023107 (2021).

[16] G. Resbiantoro, R. Setiani, and Dwikoranto, A review of misconception in physics: The diagnosis, causes, and remediation: Research article, J. Turk. Sci. Educ. **19**, 403 (2022), https://www.tused.org/index.php/tused/article/view/924.

[17] M. M. Hull, A. Jansky, and M. Hopf, Does confidence in a wrong answer imply a misconception?, Phys. Rev. Phys. Educ. Res. **18**, 020108 (2022).

[18] S. Mi, J. Ye, L. Yan, and H. Bi, Development and validation of a conceptual survey instrument to evaluate senior high school students' understanding of electrostatics, Phys. Rev. Phys. Educ. Res. **19**, 010114 (2023).

[19] J. ichiro Yasuda and M. aki Taniguchi, Validating two questions in the Force Concept Inventory with subquestions, Phys. Rev. ST Phys. Educ. Res. **9**, 010113 (2013).

[20] J. ichiro Yasuda, N. Mae, M. M. Hull, and M. aki Taniguchi, Analyzing false positives of four questions in the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010112 (2018).

[21] M. C. Wittmann and K. E. Black, Visualizing changes in pretest and post-test student responses with consistency plots, Phys. Rev. ST Phys. Educ. Res. **10**, 010114 (2014).

[22] R. Tapping, G. P. Lepage, and N. G. Holmes, Visualizing patterns in CSEM responses to assess student conceptual understanding, presented at PER Conf. 2018, Washington, DC, 10.1119/PERC.2018.PR.TAPPING.

[23] M. Rosvall and C. T. Bergstrom, Mapping change in large networks, PLoS One **5**, e8694 (2010).

[24] J. C. Brunson, ggalluvial: Layered grammar for alluvial plots, J. Open Source Softw. **5**, 2017 (2020).

[25] A. L. Traxler, T. Suda, E. Brewe, and K. Commeford, Network positions in active learning environments in physics, Phys. Rev. Phys. Educ. Res. **16**, 020129 (2020).

[26] A. Werth, J. R. Hoehn, K. Oliver, M. F. Fox, and H. Lewandowski, Instructor perspectives on the emergency transition to remote instruction of physics labs, Phys. Rev. Phys. Educ. Res. **18**, 020129 (2022).

[27] M. B. Atkinson and S. L. Bretz, Measuring changes in undergraduate chemistry students' reasoning with reaction coordinate diagrams: A longitudinal, multi-institution study, J. Chem. Educ. **98**, 1064 (2021).

[28] J. H. Doherty, J. A. Cerchiara, and M. P. Wenderoth, Undergraduate students' neurophysiological reasoning: What we learn from the attractive distractors students select, Adv. Physiol. Educ. **47**, 222 (2023).

[29] J. ichiro Yasuda, H. Uematsu, and H. Nitta, Validating a Japanese version of Force Concept Inventory (in Japanese), J. Phys. Educ. Soc. Japan **59**, 90 (2011).

[30] Revised Table I [Table I. Newtonian concepts in the revised Force Concept Inventory (form 081695r)], https://www.modelinginstruction.org/effective/evaluation-instruments/.

[31] Revised Table II [Table II for the Force Concept Inventory (revised form 081695r], https://www.modelinginstruction.org/effective/evaluation-instruments/.

[32] M. M. Hull, J. ichiro Yasuda, and N. Mae (to be published).

[33] M. J. Kolen and R. L. Brennan, *Test Equating, Scaling, and Linking* (Springer, New York, 2014), p. 2.

[34] S. L. Wise, Rapid-guessing behavior: Its identification, interpretation, and implications, Educ. Meas. **36**, 52 (2017).

[35] E. F. Redish, *Teaching Physics with the Physics Suite* (John Wiley & Sons, New York, 2003), p. 26.

[36] L. Bao, K. Hogg, and D. Zollman, Model analysis of fine structures of student models: An example with Newton's third law, Am. J. Phys. **70**, 766 (2002).

[37] T. I. Smith and M. C. Wittmann, Comparing three methods for teaching Newton's third law, Phys. Rev. ST Phys. Educ. Res. **3**, 020105 (2007).

[38] C. Chen, L. Bao, J. C. Fritchman, and H. Ma, Causal reasoning in understanding Newton's third law, Phys. Rev. Phys. Educ. Res. **17**, 010128 (2021).

[39] L. Bao and J. C. Fritchman, Knowledge integration in student learning of Newton's third law: Addressing the action-reaction language and the implied causality, Phys. Rev. Phys. Educ. Res. **17**, 020116 (2021).

[40] K. F. Wilson and D. J. Low, on second thoughts…: Changes of mind as an indication of competing knowledge structures, Am. J. Phys. **83**, 802 (2015).

[41] D. J. Low and K. F. Wilson, The role of competing knowledge structures in undermining learning: Newton's second and third laws, Am. J. Phys. **85**, 54 (2017).

[42] H. Dedic, S. Rosenfield, and N. Lasry, Are all wrong FCI answers equivalent?, AIP Conf. Proc. **1289**, 125 (2010).

[43] D. E. Brown and J. Clement, Misconceptions concerning Newton's law of action and reaction: The underestimated importance of the third law, in *Proceedings of the Second International Seminar: A Misconceptions and Educational Strategies in Science and Mechanics*, edited by J. D. Novak (Cornell University, Ithaca, 1987), Vol. 3, p. 39, https://www.google.co.jp/books/edition/Proceedings_of_the_Second_International/rWJQAAAAYAAJ.

[44] E. F. Redish, J. M. Saul, and R. N. Steinberg, On the effectiveness of active-engagement microcomputer-based laboratories, Am. J. Phys. **65**, 45 (1998).

[45] D. Hammer, Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions, J. Learn. Sci. **5**, 97 (1996).

[46] P. Nieminen, A. Savinainen, and J. Viiri, Force Concept Inventory-based multiple-choice test for investigating students' representational consistency, Phys. Rev. ST Phys. Educ. Res. **6**, 020109 (2010).

[47] S. Bogart, Sankeymatic, http://sankeymatic.com.

[48] J. ichiro Yasuda, Sample code for alluvial diagrams, https://github.com/Jun-ichiro/sample-alluvial.

[49] J. ichiro Yasuda, N. Mae, M. M. Hull, and M. aki Taniguchi, Optimizing the length of computerized adaptive testing for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **17**, 010115 (2021).

[50] J. ichiro Yasuda, M. M. Hull, and N. Mae, Improving test security and efficiency of computerized adaptive testing for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **18**, 010112 (2022).