Does confidence in a wrong answer imply a misconception?

Michael M. Hull,[1] Alexandra Jansky,[2] and Martin Hopf[1]
[1]*Austrian Educational Competence Center, Division of Physics (AECCP), University of Vienna, Vienna, Vienna 1090, Austria*
[2]*Bern University of Applied Sciences, Department Architecture, Wood and Civil Engineering, Biel 2500, Switzerland*

(Received 22 November 2021; accepted 8 June 2022; published 2 August 2022)

Our study investigates whether confidence correlates with consistency in reasoning, specifically about radioactive decay. In prior work, we developed and tested a survey designed to measure consistency of student reasoning about radioactive decay by comparing responses to three prompts that are isomorphic, meaning that, despite having different surface features, they can all be answered appropriately with the understanding that radioactive decay occurs at random. In this paper, we compare (i) student patterns on these isomorphic prompts with (ii) confidence ratings that students provided together with their responses. Our research question is "to what extent does student confidence correlate with consistency in reasoning about radioactive decay?" We have found that there is no significant correlation, suggesting that more confident students are not more likely to be consistent. One reason why this finding is relevant is that the misconceptions model attributes consistency to student ideas (as opposed to the pieces model, which describes student ideas as potentially being context dependent). Our findings suggest that it is premature to describe a student idea as a misconception, even if the student is confident in that idea.

DOI: 10.1103/PhysRevPhysEducRes.18.020108

I. INTRODUCTION

Much of physics education research concerns investigation into student ideas and how those ideas interact with what students are meant to learn in the classroom. Various theoretical models have been proposed to describe these student ideas, and it remains a question of debate which model is most effective in which situation. One point in which these models differ is the underlying assumption about the robustness and rigidity of the student ideas themselves. For example, whereas the framework theory model of Vosniadou and colleagues (e.g., Ref. [1]) attributes relative stability and difficulty to change to student ideas, the knowledge in pieces model of diSessa and others views student ideas as being potentially context dependent and fluid, shifting moment by moment (e.g., Refs. [2,3]). Scherr [4] identified eight properties of student ideas in setting up a dichotomy between the "misconceptions" and "pieces" models of student thinking, including that a misconception is context independent, stable, and difficult to change. In this paper, we use the term "misconception" to indicate a stable cognitive structure that is activated in a wide variety of contexts and that is difficult to change. Consistent with Scherr [4], we use the phrase "student idea" as a general term that includes misconceptions but also includes ideas that are transient in nature, comprised of knowledge pieces that are only loosely connected and are hence context sensitive.

Recently, education researchers have turned their attention to developing surveys which can gain insight into the structure of student ideas [5–12]. Our previous work [13–16] has utilized isomorphic problems (e.g., Ref. [17]), problems that require the same conceptual understanding to answer but have different surface features that may result in a given student answering correctly on only some of the problems. Singh argued that the reason for this fluidity in reasoning is that "problem context with distracting features can trigger the activation of knowledge that a student thinks is relevant but which is not actually applicable in that context" [17]. In such a case where knowledge is triggered in some problem contexts but not others, it is inappropriate to think of students as having a stable and resistant misconception.

Although the use of isomorphic problems is a relatively direct indicator of how context-sensitive a student's reasoning is, it has the disadvantage of requiring additional survey items, potentially increasing the length of the survey dramatically. In this paper, we consider a second tool that could potentially indicate the robustness of a learner's ideas: confidence ratings. After each item in our survey, respondents are asked how confident he or she is with the answer to that item. Hasan *et al.* have argued that

confidence ratings can "differentiate between a lack of knowledge and a misconception" [8] and Lemmer has argued that expressing confidence can "confirm the existence of stable existing [knowledge] structures" [5]. Asking respondents for a confidence rating increases the survey length only marginally. To minimize testing time, then, it would be preferable to use only confidence ratings, assuming the information they provide is equivalent to that obtained from isomorphic problems. The idea that high confidence indicates a misconception, however, remains at present a hypothesis in need of testing. Our current research contributes to doing just that. The purpose of our study is to investigate whether confidence correlates with consistency in reasoning, specifically about radioactive decay. Our research question is "to what extent does student confidence correlate with consistency across the isomorphic prompts?" In particular, is there sufficient correspondence between confidence ratings and consistency in responses to isomorphic problems that we can keep just one of the two approaches? Concretely, are students who answer a prompt incorrectly with confidence more likely to also be incorrect on the other isomorphic prompts?

A. Context for the study: Radioactivity

To date, researchers have documented a number of student ideas regarding radioactivity (e.g., Ref. [18]), including several student ideas pertaining to the time it takes for radioactive nuclei to decay [19–21]. It has been argued that part of the underlying difficulty could be a failure to understand the stochastic nature of radioactivity [18,22]. In particular, some studies examining student reasoning about radioactive decay have found that many students assume that if half of a radioactive substance has decayed after one half-life, then half of each individual atom making up the sample must have correspondingly decayed [14,23,24].[1]

We have also argued that, particularly regarding ideas about the timing of the radioactive decay of an individual atom, the difficulty could arise in part because of a failure to understand radioactivity as an emergent process [14]. Wilensky *et al.* (e.g., Ref. [25]) discussed how students demonstrate a "level confusion" when they assume that the agent level (in this case, individual nuclei) and the system level (the overall radioactive sample) share the same property (being half-decayed after one half-life). In our prior work, we have documented that this "level confusion" is not necessarily a stable and rigid cognitive structure; students often seem to have myriad ideas available to them simultaneously, including ideas about the random decay time of individual nuclei, and different ideas come to the fore depending upon the specific context [14,16].

[1]In fact, for a single atom, there is a 50% chance for radioactive decay to occur in a time period of one half-life, provided, of course, that the nucleus has not yet decayed at the start of that time period.

B. Theoretical background

A significant amount of research has documented fluidity in student reasoning through think aloud interviews and in the classroom as students respond to various contextual cues (e.g., Refs. [3,26–32]). Similarly, free response survey prompts allow for students to describe a wide range of ideas that they have pertaining to a given situation, as opposed to having to choose the one that is most salient to them, as they generally must do with multiple choice surveys. The disadvantage of free-response prompts instead of multiple choice prompts, however, is the time demanded for students to fill out the survey and time demanded for education researchers or teachers to assess the responses. A survey which is purely multiple choice has the greatest potential to reduce time demands on both respondents and on those who assess those responses. In this article, we discuss two approaches to probe for fluidity of reasoning in a multiple choice format, (i) the use of isomorphic prompts, and (ii) the use of confidence ratings.

1. Isomorphic prompts

One approach to measuring the stability of a respondent's ideas that has the capability to remain entirely in the multiple choice format is that of isomorphic prompts. The method of isomorphic prompts can be traced back to Simon and Hayes [33–35], who, in their investigation of problem solving, defined two problems as isomorphic if they had problem spaces with the same structure. As such, the similarity of isomorphic problems can be as superficial as being the same problem but with different numbers used for relevant parameters; similarly, the same problem posed twice, once asking for a numerical answer and once asking for a qualitative relationship, constitute an isomorphic pair of problems. Singh [36] wrote that two prompts are isomorphic if they "require the same physics principle to solve them." As such, one can make "the surface features of the problems very different as in the problem pair chosen by Simon and Hayes [the 'Tower of Hanoi' and 'the cannibal and the missionary problem'] or by introducing distracting features into one of the problems."

Although the isomorphic problems discussed in Ref. [36] are free response, Singh discussed in Ref. [17] other pairs of problems that are multiple choice. For example, on one item of her survey (Q.21), many students chose the correct answer "(a) $F = f$ because the mass is not accelerating" when thinking about two people pulling in opposite directions on a stationary box.[2] On an isomorphic

[2]Singh's Q.21 reads "Arnold and you are both pulling on a box of mass M that is at rest on a frictionless surface, as shown. Arnold is much stronger than you. You pull horizontally as hard as you can, with a force f, and Arnold keeps the mass from moving by pulling horizontally with a force F. Which one of the following is a correct statement about the magnitude of Arnold's force F? g is the magnitude of the acceleration due to gravity.

prompt (Q.22), however, when asked how large friction must be on a table such that the table does not move when you push it,[3] many students reasoned in a different way, saying that the friction force is equal to the coefficient of static friction times the normal force (i.e., larger than the force of the push). Even in this multiple choice format, we can see that Newton's first law is not necessarily a rigidly held idea that students will consistently use. In the presence of distracting features (such as a coefficient of friction provided in the problem statement), students may resort to formulaically setting the friction force to equal the coefficient of static friction times the normal force. This context sensitivity of student reasoning is direct evidence that the student ideas of the participants in Singh's study had not crystallized into misconceptions. Rather, they were comprised of loosely bound pieces of knowledge that could readily rearrange from isomorphic prompt to isomorphic prompt. Singh described this phenomena in terms of the pieces model of student ideas: "From the perspective of knowledge in pieces, problem context with distracting features can trigger the activation of knowledge that a student thinks is relevant but which is not actually applicable in that context. The student may feel satisfied applying the activated knowledge resource and may not look further for analogies to paired problems or other aids" [17]. Although using isomorphic prompts provides direct measure of how context-sensitive student ideas are, doing so dramatically increases the exam time, as the number of problems to solve is doubled if isomorphic pairs of problems are used, and tripled if problem triplets are used [17].

2. Confidence ratings

Confidence (belief in one's own ability) is seen as a desirable characteristic, particularly in leaders, as it inspires trust from others [37]. Someone who is confident gives the impression of being capable of succeeding in whatever she or he is confident in. At the same time, a well-known finding from social psychology is the Dunning-Kruger effect, where low-performing individuals are found to be more likely to overestimate their performance [38]. In physics education research, some studies have similarly demonstrated that poorly performing students have disproportionately high degrees of confidence (e.g., Ref. [39]). Klein *et al.* [9] compared physics and economics students when answering questions about slope and area of graphs. They found that physics students appropriately expressed less confidence when their answers were incorrect than they did when their answers were correct. Lower-performing economics students, on the other hand, were equally confident regardless of correctness. Eshach *et al.* [7] found that Taiwanese middle school students who performed poorly on a survey regarding the process property of sound were just as confident as students who performed well. This may have been caused by the poorly performing students overestimating their performance. Although confidence does not (generally) correlate with ability, it is intuitive that it would at least correlate with stability of reasoning. In particular, it is intuitive that students who are confident in a wrong answer are more likely to have a misconception in the sense of a self-consistent and stable cognitive structure. In fact, some literature in education research, which we will now discuss, has assumed this to be true. Nevertheless, it has remained a hypothesis that, to the best of our knowledge, has not been tested prior to our work.

The introduction of confidence ratings to physics education research can be attributed to Hasan *et al.* [8] who, as mentioned above, suggested their use as a means to distinguish between (i) wrong answers that arise from guessing and (ii) misconceptions, which they defined, in agreement with our definition above, as "strongly held cognitive structures." Hasan *et al.* asked confidence ratings of students responding to the Force Concept Inventory to argue that some items (the ones with low reported confidence) are being answered incorrectly because of "a lack of knowledge as opposed to the presence of a misconception." On the other hand, "misplaced certainty in the applicability of certain laws and methods to a specific question is an indicator of the existence of misconceptions." However, the idea that confidence in an incorrect answer indicates a "strongly held cognitive structure" (in contrast to a temporary alignment of knowledge pieces that happened to align in just such a way in the context of that particular survey prompt) remained an untested assumption.

The "Hasan hypothesis" has influenced subsequent work in physics education research (e.g., Refs. [5,10,12,40]). Leppavirta [40] found, in general, that students lacked confidence in their answers to items on an assessment pertaining to electromagnetism and wrote that the "relatively low confidence may suggest confusion regarding the subject of electromagnetics and lack of strong conceptual models of any kind." The items with which students *were* confident, on the other hand, were labeled "strongly held alternative conceptions" in keeping with the language of Hasan *et al.* Planinic *et al.* [10] administered a survey to students containing items pertaining to Newtonian mechanics and dc circuits. They conducted Rasch analysis to rank items by difficulty and then compared items between the two topics. They found that, for the difficult items, students who answered incorrectly were more confident in their wrong answer for mechanics than for circuits. In addition to the multiple choice selections and confidence ratings, students were asked to write down their reasoning on some items. Based upon this data, Planinic *et al.* described a difference in student cognitive structures between

[3]Singh's Q.22 reads "You are trying to slide a table across a horizontal floor. You push horizontally on the table with a force of 400 *N*. The table does not move. What is the magnitude of the frictional force the rug exerts on the table? The coefficient of static friction between the table and the rug is 0.60, and the coefficient of kinetic friction is 0:50. The table's weight is 1000 *N*."

mechanics and circuits. In mechanics, students demonstrated previously documented misconceptions about force being needed for constant velocity, etc. For circuits, the authors argued, student ideas were better described as hybrid models about some of the current turning into other forms of energy, etc. The researchers concluded that student ideas in mechanics are more misconceptionslike than ideas in circuits. Their motivation for the study was similar to that of Hasan *et al.* Whereas Hasan *et al.* had suggested that confidence ratings on a wrong answer can distinguish between "a lack of knowledge" and "strongly held cognitive structures," Planinic *et al.* wrote that confidence ratings can distinguish between "firmly held alternative ideas" and "answers [that] may be only transient responses." Although the Hasan hypothesis was formulated in 1999 [8], it has continued to influence research today, including the 2020 publication of Testa *et al.* [12], which considered overconfidence of respondents on a quantum mechanics survey. Overconfidence was calculated by conducting a Rasch analysis to scale the confidence and correctness scores to the same scale and subtracting the two (overconfidence = confidence − correctness). They situated their findings in light of those from Planinic *et al.* [10] to conclude that "students may lack strong mental models about the targeted QM topics, similarly to what happens in electromagnetism, but differently than in classical mechanics, where misconceptions are more deeply rooted." Planinic *et al.* wrote that "if one accepts that firm alternative conceptions might be recognized through incorrect answers provided with high confidence, then in this sample, the topic of Newtonian dynamics was the area most characterized by firm alternative conceptions" (emphasis ours). They made it clear that they *do* accept this hypothesis when they went on to write "[we] suggest that the degree to which students are confident in their answers may be used to rank students' alternative conceptions and identify those alternative conceptions that are significant and firmly held by students and therefore may be resistant to change."

Lemmer [5] found agreement with the Hasan hypothesis by looking across prompts and at discussions of students. Specifically, Lemmer administered a mechanics survey and observed, for example, that the majority of students believed that a kicked soccer ball that rolls and then stops "without anyone else touching it" will either get faster or travel with a constant speed after leaving contact with the foot. This, Lemmer argued, is a manifestation of the idea that "changes take time" [3]. The goal of the study was "to investigate the context dependency and stability of the changes-take-time perception…" and they found it to be relatively context independent. On a set of three problems involving rolling along a ramp (with positive, zero, and negative incline), "about 80% of the students who marked the changes-take-time option in one of these items also marked it in the other items." This, together with student insistence in their ideas during focus group discussions, was used as evidence for stability of the idea. In such a case, it is appropriate to think that the students have a misconception that things get faster after being released. Furthermore, students were confident in these answers, so Lemmer wrote that "their confidence confirms the existence of stable existing structures," consistent with the Hasan hypothesis. In many regards, our work reported here is similar to that of Lemmer. Although they were not labeled as such, the three ramp problems used by Lemmer are isomorphic at the level of physics concepts, and both consistency and confidence on these prompts were considered. However, contrary to what the Hasan hypothesis would predict, Lemmer did *not* find low confidence in cases of inconsistency. In fact, Lemmer found that "about two-thirds of the students (68%) were sure or very sure that they responded correctly… for all the questions." In our study, we will consider both cases of high *and* low confidence to see if confident students are *more* likely to be consistent in their incorrect reasoning patterns (indicating a potential misconception). In so doing, we provide the first (to the best of our knowledge) statistical test of the hypothesis that confidence in an incorrect answer indicates, in the words of Hasan *et al.*, "strongly held cognitive structures" [8]. As we will discuss in Sec. II, our test involves comparing consistency of student responses to isomorphic prompts (which provides a direct measure of how strongly held cognitive structures are) to their confidence ratings. In other words, we look to see if confidence correlates with consistency. We address our research question from two directions, with two corresponding subresearch questions:

SRQ1: "Are confident students more likely to be consistent in their answers to the isomorphic prompts?" and
SRQ2: "are students who answered the isomorphic prompts in a consistent manner more likely to be confident?"

II. METHODOLOGY

Our research is quantitative in nature, and we address our research question via analysis of multiple choice survey data. This survey data consists of isomorphic prompts and a confidence rating for each of those prompts. These confidence ratings ask students "How confident are you with your answer?" and are Likert scale, from 1 to 5, with 1 being "not at all." The survey we discuss in this paper was motivated by a pilot study of 7 interviews that utilized an approach similar to Brown and Clement's bridging strategy [41] to explore fluidity of student reasoning about half-life. To see how widespread and to what degree this context dependency is, we created the Stochastic World of Radioactive Decay Evaluation (SWORDE, pronounced "sword"). For details about these 7 interviews, the first version of SWORDE (originally named FAROS), and the

results obtained with this first version, see Refs. [14,16]. SWORDE is written in the German language, with prompts drafted by Hull and modified after discussion with Jansky and Hopf, who are both native German speakers. Quotes from survey items and responses to those items presented here are English translations that were drafted by Hull, confirmed by either Jansky or Hopf, and reconfirmed by a native German-speaking member in the AECCP at the University of Vienna (who is not himself an author).

A. Overview of the survey prompts

SWORDE consists of four prompts, (1) a Concept Cartoon [42] in which four students discuss radiation sent out by a radioactive stone while an ant is standing nearby it for 10 min (referred to as "ANT" in this paper, see Fig. 1), (2) a situation of having one's closet full of radioactive gas and wanting to retrieve valuables from inside ("CLOSET"), (3) a decision of which day would be best to watch an individual unstable atom trapped in a cage undergo radioactive decay ("CAGE") [24], and (4) the "many vs one" ("MvO") prompt, asking students to compare the amount of a radioactive substance that has not yet decayed after an integer number of half-lives, when starting with many atoms ("MvO:MANY") and when starting with just one atom ("MvO:ONE"). Each of these four prompts in the most recent version of SWORDE specifies that the radioactive substance is I-131. At the start of CLOSET, CAGE, and MvO, students are provided with an explanation of half-life that includes the information that, after 8 days, half of the I-131 sample will "have transformed into a different atom" if one begins with a large number of atoms.

In ANT, survey respondents are asked which of the four student statements they agree with, which they disagree with, and why. The prompt is multiple response, meaning that respondents can agree or disagree with more than one of the four student statements. The current version of ANT begins with the case of the stone containing "a very small" amount of I-131 ("ANT:TRACE"). Respondents are then asked how their answers would change if the stone instead had "a huge amount" of I-131 ("ANT:HUGE"). We considered at first having ANT consist of a number of true or false statements, similar to the work of Woithe [43], but decided to avoid this approach because of the intentionally ill-defined nature of the prompt. Student C, for example, says that the amount of radiation leaving the stone each moment can be calculated. Of course, it *can* be calculated; generally the issue, rather, is what degree of accuracy behind that calculation is desired. We are interested in whether this idea (of being calculable) happens to resonate with students or not (and whether it resonates more in ANT:HUGE than in ANT:TRACE), and that is subtly different from whether students think the idea is true or false. More generally, the ambiguity of what exactly a "huge amount" or a "very small amount" entails makes it unreasonable to ask which statements are true and which are false. Instead, in the current version of ANT, respondents are told to "select all statements that are consistent with your justification." Our strategy to use vague wording in this prompt is similar in some ways to that of Millar [44], whose survey prompts included options of "the [object] is slightly radioactive" and "the [object] is radioactive." Millar noted that "students have, in general, no way of deciding whether a quantity of radioactive material, or radiation, or a level of radioactivity is large or small, and these terms are, in any case, loose and open to interpretation." Millar included the selection of "slightly radioactive" only because, based upon prior findings, students are more likely to select the option when the qualifier "slightly" is attached. For analyzing the data, Millar coded responses of slightly radioactive the same as responses of "radioactive." In our case, what is relevant for our analysis is whether language about "very small" and "huge" amounts cue in students ideas about the law of large numbers and if they hence say that there would be differences in their answers or not. Unlike the other three prompts, we do not provide information about the half-life of the radioisotope in ANT, because we feared that students would become preoccupied with trying to do a time-consuming calculation. Our interest, rather, was whether or not students would respond that their answers "would not change" when going from ANT:TRACE to ANT:HUGE, which would suggest a level confusion [25] (see coding description, later).

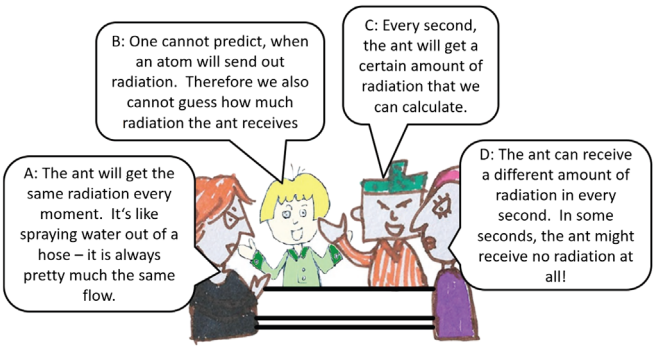

FIG. 1. The first prompt in SWORDE is ANT, where four students are discussing the radiation received by an ant standing beside a radioactive stone for 10 min, first where the stone contains "a very small" amount of radioactive substance (ANT:TRACE) and then when the stone contains "a huge amount" (ANT:HUGE).

The analysis of the free response data from CAGE and MvO has been reported in Refs. [14,16], and readers are referred to Refs. [13,15] to learn more about the development of ANT. Smaller adjustments to the multiple choice versions of the prompts are discussed here and in the Supplemental Material [45]. The development of CLOSET is discussed in Ref. [46] and, as it is not one of the isomorphic prompts, is not relevant to this article and will not be discussed further.

B. Overview of the survey creation process

We designed SWORDE to document quantitatively the context sensitivity of student reasoning about radioactive decay that we first observed qualitatively in interviews. To this end, we have completed five cycles of revision, data collection, and analysis of results (to inform the next set of revisions to the survey). The first version of SWORDE, consisting of only free response prompts, was administered to $N = 55$ 14-yr-old students in 2019. Three survey validation interviews were conducted prior to this first administration (SVI#1, SVI#2, and SVI#3), in addition to the seven semistructured interviews that served as the foundation for SWORDE. For additional details about the first version of SWORDE and the results obtained with this survey, see Refs. [14,16].

The second version of SWORDE consisted of just MvO and CAGE in a closed (multiple choice) form that Hull administered to his preservice teachers at the start of a seminar on radioactivity. The multiple choice forms of these prompts were made directly from the qualitative content analysis [47] that was done with the free response data from the $N = 55$ respondents who answered version 1 (see Refs. [14,16]).

The biggest changes on version 3 of SWORDE were specific to ANT, which remained in a free response form [13]. Smaller changes included having the language consistently talk about atoms "transforming" instead of "remaining" to avoid reinforcing the misconception about atoms disappearing when they decay [21]. Three additional survey validation interviews (SVI#4, SVI#5, and SVI#6) were conducted before releasing this third version of SWORDE. This third version of SWORDE (multiple choice for all prompts except ANT) was released in early June 2020 to $N = 37$ 18-yr-old students who had already learned about radioactivity and half-life (and who, when asked on the survey if they had learned about half-life previously, did not deny it).

In preparing for the fourth version of SWORDE, the free response version of ANT was replaced with a multiple choice version [13] and confidence ratings were introduced to the survey for the first time. An additional survey validation interview (SVI#7) was carried out before the fourth version of SWORDE was administered. This fourth version of SWORDE was administered two weeks after version 3 was administered, this time to $N = 47$ 18-yr-old students who had already learned about half-life and radioactivity (and who, when asked on the survey if they had learned about half-life previously, did not deny it).

The biggest change made prior to the release of version 5 of the survey (the final version) was the removal of the "other" options from the multiple choice prompts. This decision was justified by careful examination of the 84 student responses on versions 3 and 4 of SWORDE and two additional survey validation interviews (SVI#8 and SVI#9) of students chosen from that combined dataset (see Supplemental Material [45] for details).

Throughout this process, online survey validation interviews (SVI#4-9) were conducted by Hull via Zoom. Students who took place in a survey validation interview were compensated with a certificate of completion and a tree planted in their name through TreeNation. The fifth (final) version of SWORDE was then administered online starting November 2020 via Survey Monkey. Whereas respondents to the first four versions of SWORDE were selected mostly out of convenience (teachers of schools with which the University of Vienna already had connections were personally invited to participate), respondents to the fifth (final) version were a more random sample. Specifically, physics teachers all across Austria were invited to participate via a mailing list, with the specification that the survey was intended for students in the 11th and 12th grades (17–18 yr olds).

C. Survey administration and analysis

In between November 9, 2020 and March 24, 2021, SWORDE collected data from 527 respondents. Of these, 80 respondents were instructors who were considering administration of the survey to their students. The survey was administered online and an answer was mandatory to move on to the subsequent question. Since our research question involves looking for consistency of student responses across three isomorphic prompts, and since the third isomorphic prompt was the final item of the survey, we removed all survey respondents (157) who did not complete the survey. The survey begins by asking students if they have learned about half-life before. As mentioned above, instructors were told that the survey was intended only for students in 11th and 12th grades. Although the topic of half-life is part of the national mathematics curriculum in 10th grade, 24 of the remaining students answered that they had not learned about half-life previously. Their responses were removed. Finally, the first part of MvO:MANY (asking for how much of the radioactive substance will remain after one half-life) served as a screening question to remove an additional 32 respondents. Specifically, despite the explanation about half-life just prior to the prompt, a number of students nevertheless selected either "100 million atoms" (the starting amount), "0 atoms," or "100 million OR 0 atoms" for the amount that would have not yet decayed after 8 days. Since we had encountered no difficulty in understanding this prompt in any of the nine survey validation interviews that were involved with the survey creation, we assumed that these responses were due to random guessing and we accordingly removed the respondents from our dataset. After these measures were taken, a total of $N = 234$ student responses remained, and these responses were used for all analyses discussed in this article. We justified grouping this data into one dataset, despite being collected from multiple classes, based upon results from a cluster analysis [48].

Specifically, although 3 clusters were identified, these clusters did not coincide with classes of students.[4]

1. Coding of responses

The items we coded consist of confidence ratings and isomorphic prompts. Although the confidence rating data was collected on a five-point Likert scale, we collapsed the data into three levels so as to increase the number of respondents in each category. This is a practice that has been done by many other studies in physics education research using Likert-scale data (e.g., Refs. [49–51]), with the justification that many students are inconsistent in their distinguishing between, for example, "strongly agree" and "agree." As such, we assigned a "1" to students who were "not at all confident" or "not confident" on an item, a "2" to students who selected "neutral," and a "3" to students who reported being "confident" or "very confident."

Based upon the responses to the isomorphic prompts, we looked for consistency of student reasoning. In particular, we are interested in two types of consistency in responses: (i) consistently indicating the view that what is true for the radioactive sample is also true for the individual nucleus (that is, a level confusion [25]); (ii) consistently indicating the understanding that radioactive decay is a process that occurs at a random point in time. For each of these, checking for consistency required coding at two levels. First, we assigned a level confusion (LC) code to each response that indicated a level confusion and a "randomness" (RAND) code to each response that demonstrated awareness of randomness being relevant for answering the survey item (see next section: "Coding each response to each prompt"). We then coded each respondent as being consistent (or not) in receiving these codes across the three isomorphic prompts (see "coding each respondent").

Coding each response to each prompt.—On ANT:TRACE, respondents were asked to consider the case when the stone contains a very small amount of I-131. We assigned a RAND code if the respondent agreed with student B: "One cannot predict, when an atom will send out radiation. Therefore we also cannot guess how much radiation the ant receives" (first row of Table I). We also assigned a RAND code if the respondent agreed with student D: "The ant can receive a different amount of radiation in every second. In some seconds, the ant might receive no radiation at all!" (second row of Table I). In total, we assigned at least 1 RAND code to 104 students (third row of Table I). On ANT, there are many possible responses that one can interpret as indicating a level confusion [13]. We decided, however, to be consistent with our prior work [15] and only assign a LC code when the clearest indication of a level confusion is given. Therefore, we assigned a LC code on ANT:HUGE if the respondent (1) selected the option "my answers would not change" and/or (2) selected "nothing changes except that the stone now sends radiation out longer." From these criteria alone, we assigned a LC code to 196 of the 234 respondents.

On CAGE, respondents are told that the half-life of I-131 is 8 days and they are asked what day they would go to watch a single I-131 atom decay and why (multiple select). Here, we focused exclusively on the second tier (the reasoning tier) for this prompt (see Supplemental Material [45] for justification). We assigned a RAND code if the respondent selected "it is unpredictable when the atom transforms" for the reasoning tier. We assigned a LC code to responses of "the atom transforms continuously," "after the half-life, half of the atom will have transformed," and "the atom transforms on the day of the half-life" on the reasoning tier.

On MvO, respondents are asked how much I-131 would have not yet decayed if one begins with 100 million atoms (MvO:MANY) and if one begins with just one atom (MvO:ONE). For both parts, students are also asked to provide one or more reasons. Evidence for understanding of randomness can be found in MvO:ONE. We assigned a RAND code to the response of "one atom OR no atoms" (have not yet transformed) on each part (8, 16, and 24 days) of the answer tier. We also assigned a RAND code to the response of "it is random" on the reasoning tier. We assigned a LC code to the response of "one half atom" on the answer tier for 8 days. We also assigned a LC code for the response of "after the half-life, half of the atom will have transformed" on the reasoning tier. Finally, with a justification readers can find in the Supplemental Material [45], we assigned a LC code if a respondent correctly chose "50, 25, 12.5" as the answers to MvO:MANY but then chose either "1, 0, 0" or "0, 0, 0" as the answers to MvO:ONE with the reason of "One cannot have half an atom."

Coding each respondent.—We coded for context dependency in the form of consistency across the prompts to ask "does a respondent consistently indicate a level confusion on all three isomorphic prompts?" Our outcome variable is hence "consistently demonstrates a level confusion," abbreviated "LCC" (with the last "C" standing for "consistent"), as measured by receiving at least one LC code on each of the three isomorphic prompts (i.e., LCC:yes) or not (i.e., LCC:no). Note that this label is binary, so students who receive a LC code on 2, 1, or 0 isomorphic prompts have the LCC variable set to "no."

[4]Since SWORDE was administered online, it was possible for students to take the survey on their own outside of regularly scheduled class time. To give maximum confidentiality to student responses, no demographic information, including teacher name or class, was recorded. However, by looking at the starting time of the survey responses, we were able to find students who may have been in the same class as each other. Specifically, before carrying out the data cleaning process described above, we arranged all respondents by starting time of the survey. Respondents who started within 5 min of each other during typical school hours were assumed to be within the same class.

TABLE I. Codes from the three isomorphic prompts, $N = 234$. We assigned LC codes to responses that indicated a level confusion and RAND codes to responses that indicated awareness of the randomness of radioactive decay.

Prompt	Code	Reason	*N*
ANT	RAND	TRACE: Agree w/B: "One cannot predict…"	32
		TRACE: Agree w/D: "… can receive a different amount…"	91
		TOTAL	104[a]
	LC	ANT: HUGE: "My answers would not change"	99
		ANT: HUGE: "Nothing changes except that the stone now sends radiation out longer"	127
		TOTAL	196
CAGE	RAND	"It is unpredictable when the atom transforms"	59
		TOTAL	59
	LC	"The atom transforms continuously"	51
		"After the half-life, half of the atom will have transformed"	96
		"The atom transforms on the day of the half-life"	44
		TOTAL	154
MvO	RAND	ONE, 8 days: "One atom OR no atoms"	73
		ONE, 16 days: "One atom OR no atoms"	63
		ONE, 24 days: "One atom OR no atoms"	63
		ONE: "It is random"	49
		TOTAL	96
	LC	ONE, 8 days: "One half atom"	114
		ONE: "After the half-life, half of an atom will have transformed"	114
		MANY: "50, 25, 12.5" +ONE: "1, 0, 0" +ONE: "One cannot have half an atom"	1
		MANY: "50, 25, 12.5" +ONE: "0, 0, 0" +ONE: "One cannot have half an atom"	9
		TOTAL	143

[a]Multiple RAND and LC codes were possible on a given prompt, but we did not give additional weight than if the student received just one such code. The "Total" rows are therefore not the sum of the numbers of individual codes. For example, a total of 104 students received one or two RAND codes on ANT.

TABLE II. Three example survey responses (from $N = 234$) for considering correlation between confidence and consistency in LC codes[a]. "Resp." is an abbreviation for "Respondent" and "Conf." is an abbreviation for "Confidence".

	LC codes to isomorphic prompts and confidence						
Resp.	ANT Code	CAGE Code	MvO Code	LCC Code	Conf. ANT	Conf. CAGE	Conf. MvO
1	×	LC ("After the half-life, half of the atom has transformed")	LC (MANY: "50, 25, 12.5" +ONE: "0, 0, 0" +ONE: "One cannot have half an atom")	no	3	1	2
2	LC ("My answers would not change")	×	×	no	2	1	2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
6	LC ("My answers would not change")	LC ("After the half-life, half of the atom has transformed")	LC (ONE: "After the half-life, half of an atom will have transformed")	yes	1	1	2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
234	…	…	…	…	…	…	…

[a]LCC: yes students received the *LC* code on all three prompts. A "1" in the right-most columns indicates not at all confident or not confident, a "2" indicates neutral, and a "3" indicates very confident or confident.

Let us consider a few examples of survey responses while referring to Table II. To answer our research question, we look for correlation between consistency across the three prompts and confidence on a given prompt on a prompt-by-prompt basis. For example, are students who received a LC code on ANT and are confident in that response more likely to be coded LCC:yes? To answer this question, we exclude respondents who did not receive LC codes on ANT (like respondent 1 in Table II). Respondent 2, on the other hand, did receive a LC code on ANT, and so we include this respondent. Since we did not assign a LC code to respondent 2 for CAGE and MvO, we coded this respondent as LCC:no. On ANT:HUGE, respondent 2 reported a confidence of "neutral," which we assign a "2." Hence, there is a "2" in the "Conf. ANT" column for respondent 2 in Table II. As a final example, we assigned a LC code to respondent 6 not only on ANT, but on all three prompts, and hence coded the respondent as LCC:yes. On ANT:HUGE, this respondent reported a lack of confidence (either "not at all confident" or "not confident"), which we coded as "1." Hence, next to respondent 6 in the table, in the "Conf. ANT" column, a "1" has been entered. When looking at confidence ratings on CAGE, we included respondent 1 (LCC:no, Conf. CAGE:1) and respondent 6 (LCC:yes, Conf. CAGE:1), but not respondent 2. When looking at confidence ratings on MvO, we again included respondent 1 (LCC:no, Conf. MvO:2) and respondent 6 (LCC:yes, Conf. MvO:2), but not respondent 2.

In the same way, with a similar argument for students who demonstrated awareness of radioactive decay as a random occurrence, we divided students into two groups: RANDC:yes and RANDC:no (where the last C again stands for "consistent") and used a table analogous to Table II to organize our data. From these confidence ratings, LCC codes, and RANDC codes, we conducted logistic regression and the Mann-Whitney U test to address our research question.

III. RESULTS

In Table II, we showed three examples of the LCC codes assigned. Table III presents a summary of all $N = 234$ respondents. Respondent 2 in Table II, for example, is one of the 25 LCC:no respondents with a confidence of "2" in the left columns (pertaining to ANT) of Table III. Respondent 6 in Table II is one of the 51 LCC:yes respondents with a confidence of "1" in the left columns of Table III. Respondent 6 is also one of the 55 LCC:yes respondents with a confidence of "1" in the middle columns (pertaining to CAGE) and one of the 23 LCC:yes respondents with a confidence of "2" in the right columns (pertaining to MvO) in Table III.

Overall, we can see that *for both confident and unconfident students,* there are more LCC:yes than LCC:no students on both CAGE and MvO. To investigate our research question more thoroughly, we turn to our two subresearch questions.

TABLE III. Confidence level of students who were assigned a level confusion code on ANT (left), CAGE (middle), and/or MvO (right). Out of $N = 234$ respondents, 102 received a LC code on all three isomorphic prompts. These are the respondents with LCC:yes.

	LC on ANT			LC on CAGE			LC on MvO		
	LCC			LCC			LCC		
Confidence	*no*	*yes*	Total	*no*	*yes*	Total	*no*	*yes*	Total
3 (confident)	22	16	38	9	23	32	4	23	27
2 (neutral)	25	35	60	8	24	32	13	23	36
1 (unconfident)	47	51	98	35	55	90	24	56	80
Total	94	102	196	52	102	154	41	102	143

A. Influence of confidence on consistency (SRQ1)

We will now address SRQ1: "Are confident students more likely to be consistent in their answers to the isomorphic prompts?" To do this, we focus on Table III and ask to what extent the ratio of LCC:yes to LCC:no changes with increasing confidence. On MvO, for example, the odds of an unconfident student being LCC:yes are 56:24. The odds are much better for a confident student, at 23:4, which is consistent with the idea of confident students being more likely to be consistent in their responses. We will now investigate whether these increased odds are statistically significant or not.

We are interested in seeing if the dichotomous variable (LCC:no or yes) depends upon confidence. The confidence rating items are on a 5-point Likert scale labeled at the end points with "1. Not at all" and "5. Very," but as discussed above, we collapsed this into three levels. We avoided treating the Likert scale as interval data (where the points on the scale would be considered uniformly spaced, making it possible to calculate average scores). We did this because there is controversy in the literature about whether it is appropriate to treat Likert-scale data as interval data (e.g., Refs. [52,53]). Instead, we treated the Likert scale as ordinal data, where the options are discrete categories. For these conditions, logistic regression is an appropriate choice. Zhang *et al.* [54] used logistic regression to investigate which factors predict whether or not engineering majors stay on to graduate. Since they considered multiple independent variables, they utilized multiple logistic regression. In our case, confidence is the only independent variable we consider, and so we use simple (univariate) logistic regression. Zwolak *et al.* [55] used both simple logistic regression and multiple logistic regression to see if ethnicity and other categorical variables predicted student persistence in taking a subsequent physics course. Fuad *et al.* [56] also conducted both simple logistic regression and multiple logistic regression to see which factors lead to depression in medical students.

We found coefficients in the regression by taking natural logarithms of odds ratios of values in Table III. As an

TABLE IV. Estimated coefficients and associated p values (in parentheses) for the logistic regression model for consistently demonstrating a level confusion as predicted by confidence on ANT:HUGE, CAGE, and MvO:ONE. No p values were significant.

Estimates	ANT	CAGE	MvO
β_2	0.255 (0.441)	0.647 (0.162)	−0.277 (0.514)
β_3	−0.400 (0.300)	0.486 (0.279)	0.902 (0.129)

$***p < 0.001$ $**p < 0.01$ $*p < 0.05$.

example, consider the results from ANT (left columns of Table III). For a confidence of 1 (not at all confident or not confident), for example, the odds of being LCC:yes are 51/47 (1.09). If confident students are more likely to be consistent, then we should expect these odds to be greater for students with a confidence of 3 (confident or very confident). Looking at Table III, however, we see that the odds are only 16/22 (0.73). Logistic regression considers these differences in odds by taking the logarithm of the ratio of the odds. The coefficient in the model corresponding to going from a confidence of 1 to a confidence of 3 is hence $\beta_{\mathrm{ANT},3} = \ln[(16/22)/(51/47)] = -0.400$ (model coefficients are frequently labeled as β in logistic regression). The fact that the log odds ratio is negative comes from the fact that the odds of being consistent are smaller for confident students than for unconfident students. This is directly contrary to what we would expect if confident students are more likely to be consistent in their responses. This and other relevant coefficients are in Table IV.

We are now at a point in which we can pose our subresearch question in a form appropriate for logistic regression and a corresponding null hypothesis H0 that our methods will test:

SRQ1: *Are logistic regression coefficients (log odds ratios) positive with significantly small p values?*

(H1, 1) The logistic regression coefficients are positive and the p values are significantly small.
(H1, 0) The logistic regression coefficients are either negative or the p values are not significantly small.

These coefficients and associated p values were calculated using the glm command in R Studio with the "family" set to "binomial" (since the dependent variable is dichotomous). In logistic regression, p values are found analogously to how they are found in linear regression, but instead of a continuous dependent variable (like score on an exam), the natural log of the odds (in this case, the odds of being LCC:yes) is used.

Considering the p values, we find that all results are statistically insignificant. Therefore, we *do not reject* the *null* hypothesis H1, 0. The implication of this is that, despite increased confidence, it is not necessarily the case that a student is more likely to answer consistently with *LC* responses. This conclusion was unaffected when we tried keeping all five levels of the Likert scale. Nor did it change when we relaxed our criterion for "being consistent" so that respondents who received an *LC* code on three *or two* of the isomorphic prompts were labeled as LCC:yes.

We now turn our attention to using the RAND codes we assigned to look for consistency in recognizing that radioactive decay occurs at random. We assigned the code RANDC = 1 to respondents who received at least one RAND code on each of the three isomorphic prompts. In Table V, confidence for ANT is from ANT:TRACE.

A total of 104 respondents indicated awareness of randomness on ANT, and only 23 of them also were coded with the RAND label on CAGE and MvO as well. The pattern is similar to Table III above for looking at the LC codes. We see again in Table V that about half of the respondents expressed a lack of confidence. This time, however, the consistent group of 23 respondents comprises the minority. On all three prompts, at any confidence level, there are more RANDC:no than RANDC:yes respondents. Again, however, to address SRQ1, we look not at the odds themselves, but at how the odds change as confidence increases. Unlike with the LC codes, we can see with the RAND codes that the odds change as we would expect them to. Looking at ANT, for example, the odds of being RANDC:yes with a low confidence are $(9/42 = 0.21)$ and with a high confidence $(7/13 = 0.54)$. To investigate whether these trends are statistically significant or not,

TABLE V. Confidence level of students who were assigned a RAND code on ANT (left), CAGE (middle), and/or MvO (right). Out of $N = 234$ respondents, 23 received a RAND code on all three isomorphic prompts. These are the respondents with RANDC:yes.

	RAND on ANT			RAND on CAGE			RAND on MvO		
	RANDC			RANDC			RANDC		
Confidence	no	yes	Total	no	yes	Total	no	yes	Total
3 (confident)	13	7	20	9	9	18	16	11	27
2 (neutral)	26	7	33	6	5	11	12	5	17
1 (unconfident)	42	9	51	21	9	30	45	7	52
Total	81	23	104	36	23	59	73	23	96

TABLE VI. Estimated coefficients and associated p values (in parentheses) for the logistic regression model for RANDC: yes as predicted by confidence on ANT:TRACE, CAGE, and MvO: ONE. Only one p value was significant (marked with asterisks).

Estimates	ANT	CAGE	MvO
β_2	0.228 (0.685)	0.665 (0.359)	0.985 (0.141)
β_3	0.921 (0.122)	0.847 (0.170)	1.486 (0.008)**

$***p < 0.001$ $**p < 0.01$ $*p < 0.05$.

we turn again to logistic regression. The coefficients from logistic regression are given in Table VI.

Unlike the test with the LC codes, the RANDC codes show all three prompts having positive log odds ratios (β_2 and β_3 values). This positive value is due to an odds ratio greater than 1. This in turn is due, again, to the fact that the odds of being consistent in demonstrating awareness of the randomness of radioactive decay (RANDC: yes) when the confidence is 2 (neutral) or 3 (confident" or very confident) is greater than when the confidence is 1. This, however, is only statistically significant for MvO. The effect size is the value of the estimate itself. Among students who received a RAND code on MvO and were confident in their answers, the odds ratio for being RANDC: yes in comparison to students who were unconfident was $(11/16)/(7/45)$; hence, the effect size is $\ln[(11/16)/(7/45)] = 1.486$. This value and the associated p value are relevant for SRQ1 and we find that, in the case of MvO, there is a less than 0.8% chance that the null hypothesis is correct. For the other two prompts, however, we do not reject the null hypothesis. Overall, we find mixed findings for the hypothesis H1, 1: *the logistic regression coefficients are positive and the p values are significantly small*.

B. Influence of consistency on confidence (SRQ2)

So far we have addressed our research question by focusing on the first subresearch question. We have found mixed results regarding confidence in a RAND response predicting consistency in answering with RAND responses. Regarding level confusions and the intuitive hypothesis that confidence in a LC response predicts consistency in answering with LC responses, we have not rejected the null hypothesis. We will now examine our second subresearch question, flipping our previous analysis on its head to see if consistent students are more likely to be confident.

For this subresearch question, we conceptualize our data as consisting of responses from two independent samples. In the case of LC codes, the two samples are "students who received the code LCC: yes" and "students who received the code LCC: no." We have one dependent variable (confidence rating) on each of the three items (ANT, CAGE, MvO), which is ordinal (not interval) data. Since the dependent variable is no longer dichotomous, logistic regression is no longer an appropriate analysis technique. Because ordinal data such as ours cannot be averaged, we used the two-group Mann-Whitney-Wilcoxon rank-sum test, also known as the Mann Whitney U test (U test). Use of the U test requires no assumptions about the shape of distribution, unlike the t test, which requires data from a normal distribution. The U test can be used so long as there are at least 20 data points and the data are independent. Dare and Roehrig [57] used this test to compare physics-related perceptions of girls and boys in the 6th grade. McPadden and Brewe [58] used the U test to find a significant difference in representation usage by students who had completed a semester of Modeling Instruction and those who had instead completed a semester of traditional physics instruction. Wilcox and Lewandowski [51] used the U test to compare student responses on the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS). Students were categorized as either being in a lab class that focused specifically on developing lab skills or as being in a lab class that focused on reinforcing concepts.

We again operationalize our subresearch question in a form appropriate for the U test with a corresponding null hypothesis H0 that our methods will test:

SRQ2: *Are median confidence ratings for consistent students greater than for inconsistent students (with significantly small p values)?*

> *H2, 1: The median confidence ratings are greater for consistent students and the p values are significantly small.*
> *H2, 0: The median ratings are either smaller for consistent students or the p values are not significantly small.*

The wilcox.test command was used in R Studio to calculate a p value for each of the three isomorphic prompts. If the p value is below the usually agreed-upon value of 5 percent (0.05) (e.g., Ref. [51]), the null hypothesis can be rejected and a significant difference can be assumed, as there is less than a 5% chance of obtaining these results for data collected from two identical samples. When the result of the test was statistically significant, we also determined effect size by calculating the r value from the Z value which was obtained via the qnorm command.

Regarding the LCC codes, for none of the three isomorphic prompts was the p value less than the critical value of 0.05 (see Table VII). As such, we do not reject the null hypothesis H2,0. Regarding the RANDC codes, only on MvO was the p value less than the critical value of 0.05 (see Table VIII). Hence, although we do not reject the null hypothesis in general, we can do so in the case of the MvO prompt. Therefore, as was the case with our logistic regression analysis, we found mixed results regarding the RANDC codes. The effect size was found by converting the Z value into the r value by dividing the absolute value of

TABLE VII. Results of the two-group Mann-Whitney U test and median values for each group of students who exhibited a level confusion on a given prompt.

Item	$M_{\mathrm{LCC:yes}}$	$M_{\mathrm{LCC:no}}$	U	p	r
1. Confidence on ANT:HUGE	1.5	1.5	4609	0.612	N/A
2. Confidence on CAGE	1	1	2991	0.144	N/A
3. Confidence on MvO:ONE	1	1	2270.5	0.372	N/A

$***p < 0.001$ $**p < 0.01$ $*p < 0.05$.

TABLE VIII. Results of the two-group Mann-Whitney U test and median values for each group of students who indicated awareness of radioactive decay as a random occurrence on a given prompt.

Item	$M_{\mathrm{RANDC:yes}}$	$M_{\mathrm{RANDC:no}}$	U	p	r
1. Confidence on ANT:TRACE	2	1	1095.5	0.163	N/A
2. Confidence on CAGE	2	1	498	0.155	N/A
3. Confidence on MvO:ONE	2	1	1127.5	0.006**	0.280

$***p < 0.001$ $**p < 0.01$ $*p < 0.05$.

the former by the square root of the number of observances ($n = 96$ respondents received a RAND code on MvO). The r value is less than 0.3 and so represents a small effect size [59].

C. Summary of findings

On the U test, for students who consistently indicated awareness that radioactive decay is a randomly occurring process (and so were coded RANDC: yes), MvO had a p-value less than 0.05. That is, students who answered the other two questions with a RAND response as well were more likely to express confidence in their RAND response to MvO: ONE in comparison to inconsistent students who happened to answer MvO with a RAND response. From logistic regression, we see a comparable result, that students who are more confident in their RAND response to MvO were more likely to be in the RANDC: yes group. On the other two prompts (CAGE and ANT), however, we find no correlation between consistency in RAND codes and confidence.

In looking at survey responses that indicate the student idea that what is true for the radioactive sample is true for the unstable nucleus (that is, a level confusion), we find no statistically significant evidence to support H1 or H2.

IV. DISCUSSION

In this paper, we have investigated the extent to which student confidence correlates with consistency in reasoning, specifically about radioactive decay. We explored this via the Stochastic World of Radioactive Decay Evaluation. SWORDE consists of three isomorphic prompts, which can all be answered appropriately with the same underlying physics principle; namely, that it is considered random when radioactive decay takes place. All three prompts also allow for responses that indicate the student idea that what is true for the radioactive sample is true also for the individual nucleus. Education research has found that many students think that after one half-life, since half of the radioactive sample has decayed, it follows that half of each atom in that sample has decayed [14,21,23]. This is an example of a level confusion, mistakenly assuming that what is true at the emergent level (here, the radioactive sample) is true also at the agent level (the individual atom) [25].

A total of $N = 234$ responses from 17–18-yr olds survived our data cleaning process (students who said they had not yet learned about half-life, for example, were removed). On one of the three isomorphic prompts (MvO: ONE), 114 (49%) said that, if you begin with a single atom, then half of that atom will have transformed after one half-life has passed. Across all three isomorphic prompts, 224 (96%) showed evidence of a level confusion on at least one prompt. Clearly, the idea of "what is true for the radioactive sample is true for an individual unstable nucleus" is salient for many students. However, we found that only about half of these respondents (102 respondents, 44%) exhibited a level confusion consistently across all three prompts.

We have found that, in the case of the three isomorphic prompts that we used and the coding of responses indicating a level confusion, increased confidence in a LC response does *not*, generally, indicate less context dependency of student ideas. In terms of properties associated with misconceptions (e.g., Refs. [4,60]), context-dependency across the isomorphic prompts indicates that the student idea of a level confusion has not yet crystallized into a stable structure. This is relevant as it contrasts with the writings of Hasan *et al.*, which suggested that the use of confidence ratings can distinguish between (i) wrong answers that arise from guessing and (ii) misconceptions. Specifically, they wrote that, on an individual item, "misplaced certainty in the applicability of certain laws and methods to a specific question is an indicator of the existence of misconceptions [8]." In the case of our study,

"applicability of certain laws and methods" corresponds to modeling the decay of a single unstable nucleus in terms of half-life (for example, that it is half-gone after one half-life). This hypothesis of Hasan *et al.* has influenced the work of other physics education researchers (e.g., Refs. [10–12,40,61]), and the hypothesis has been accepted in some of that work (e.g., Refs. [10,12,40]). To the best of our knowledge, however, until our study reported in this paper, the Hasan hypothesis had not been tested in the sense of seeing if students who are more confident in an incorrect answer are actually more likely to have a misconception or not. Although Lemmer *et al.* found that students with robust cognitive structures expressed confidence in those answers, they did not compare with students of lower confidence. Whereas they wrote that the majority of students "were sure or very sure that they responded correctly… for all the questions," that was not the case in our study. We found a distribution in confidence, with most students being unconfident. This allowed us to see how patterns change as confidence increases.

The answer to our research question "to what extent does student confidence correlate with consistency across the isomorphic prompts?" is "there is no statistically significant correlation." Although our study was specific to three prompts regarding student understanding of half-life and radioactive decay as a random occurrence, we suspect our findings are relevant to researchers investigating student ideas in other fields as well. Our findings suggest that, although isomorphic prompts require additional testing time, asking for confidence ratings is not an effective substitute, despite saving time. Additional studies similar to ours should be conducted to see if this is the case only with the topic of half-life, or only with these three prompts, and if confidence ratings can be used instead of isomorphic prompts in other assessments.

IV. LIMITATIONS AND FUTURE WORK

Although we have conducted a total of nine survey validation interviews throughout the creation process of SWORDE, we suspect that there is future room for improvement in terms of survey design and in terms of our interpretation of student responses. On MvO:ONE, for example, a small number of respondents said that, after 8 days, there would be 65.5 atoms left of I-131 (despite beginning with only 1 atom). Since none of the survey validation interviews included students who made this selection, we can only speculate the cause for this response. Of course, it is possible that the selection was made at random, but this option was included as a multiple choice option because a number of students had written that answer in themselves on the free response version of the survey. Another possibility is that the respondent was ardently looking for something to divide by 2, and chose the atomic mass as the only available option. An alternative interpretation, however, is that the student was aware that the I-131 nucleus contains 131 nucleons, and, since the atom is half-gone after 8 days, 65.5 of those nucleons would remain. In other words, it may be that this selection should be coded equivalently to students who selected "half an atom" as their answer. At any rate, although assigning a LC code to the response of "65.5 atoms" increased the number of students receiving at least one LC code on MvO, this did not change our results in any significant way.

Of the three isomorphic prompts we used in our study, the one that warrants the most grounds for concern is ANT. Two students (out of $N = 234$) volunteered comments at the end of SWORDE showing an understanding of ANT, despite having selected on ANT:HUGE that their answers would not change (receiving a code of LC perhaps erroneously). The first of these two respondents wrote "In the question with the ant it would have to be better defined, what trace and huge amounts of I-131 are. Because if trace means just a few atoms, then that is different than when the I-131 accounts for 1% of the mass of the stone." This respondent had selected "my answers would not change" for ANT:HUGE, presumably under the assumption that "very small amount" had nevertheless referred to the case where a sufficient amount of the stone was comprised of I-131 for the concept of half-life to still be useful for making predictions about the amount of radiation emitted, etc. The second respondent wrote "since the transformation of radioactive atoms is random, it is only probable, that radioactive iodine transforms, it could also theoretically be the case, that no radiation comes out of the stone for a short amount of time. Furthermore, 10 minutes is a short time, so radiation for the ant doesn't actually change in any noticeable way." This student selected for ANT:HUGE that "nothing changes except that the stone now sends radiation out longer." Neither of these respondents received RAND codes on ANT, although they did receive RAND codes on CAGE and MvO. Based upon their written comments just discussed, we acknowledge the absence of a RAND code on ANT as a false negative. It is also possible that the assigning of LC codes to these students was a false positive. It is likely that other false negatives and positives exist from students who, unlike those two particularly motivated respondents, did not take time to describe their understanding in written form at the end of the survey. Furthermore, out of the 80 teachers who took the survey (to preview the survey to decide whether or not to have their students complete it), seven of them expressed concern about the survey, and four of those seven expressed concern specifically about ANT. Finally, one can argue that CAGE and MvO are "more isomorphic" than ANT, since ANT does not explicitly discuss the situation of a single unstable nucleus. In view of these concerns, we tried removing ANT and using only CAGE and MvO to code students. Specifically, we assigned an LCC : yes code to respondents who received at least one LC code on CAGE as well as at least one LC code on MvO. This increased the

number of students with an LCC:yes code to 119 (instead of 102) but did not affect our conclusions.

A. Future work

As explained above, we deliberately asked no demographic information to maximize anonymity of survey respondents. Based upon our findings that increased confidence does not imply greater likelihood to respond consistently, we think it is important that we add demographic questions in to future versions of the survey, so that we can see what differences in confidence of our respondents do correlate with. For example, Leppavirta [40] found that, in comparison to male classmates, female students expressed less confidence in their answers on a survey about electromagnetism. As such, we intend to ask about respondent gender in future versions of the survey to see if differences in confidence correspond better to gender than to consistency in reasoning across the isomorphic prompts. This and other demographic questions would be placed at the end of the survey to prevent the questions from affecting responses on subsequent physics questions due to stereotype threat (e.g., Ref. [62]). Interest in physical science in general and radioactivity in particular is an additional factor that may influence confidence, and future studies should investigate this as well.

ACKNOWLEDGMENTS

The authors wish to thank Shizuka Nakayama for the illustrations that accompany our papers and presentations. We also wish to thank Marko Lüftenegger for valuable advice regarding the statistics we employed. We acknowledge Axel-Thilo Prokop, whose research findings on the understanding of preservice teachers about radioactivity inspired the current name of our survey.

[1] S. Vosniadou and I. Skopeliti, Conceptual change from the framework theory side of the fence, Sci. Educ. **23**, 1427 (2014).
[2] A. A. diSessa, A Bird's-Eye View of the "Pieces" vs. "Coherence" Controversy (From the "Pieces" Side of the Fence), in *International Handbook of Research on Conceptual Change*, edited by S. Vosniadou (Routledge, London, 2009), p. 35.
[3] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).
[4] R. E. Scherr, Modeling student thinking: An example from special relativity, Am. J. Phys. **75**, 272 (2007).
[5] M. Lemmer, Nature, cause and effect of students' intuitive conceptions regarding changes in velocity, Int. J. Sci. Educ. **35**, 239 (2013).
[6] N. K. Weinlader, E. Kuo, B. M. Rottman, and T. J. Nokes-Malach, *A new approach for understanding student resources with multiple-choice questions, presented atPER Conf. 2019, Provo, UT* (AIP, Melville, NY, 2019), 10.1119/perc.2019.pr.Weinlader.
[7] H. Eshach, T. C. Lin, and C. C. Tsai, Taiwanese middle school students' materialistic concepts of sound, Phys. Rev. Phys. Educ. Res. **12**, 010119 (2016).
[8] S. Hasan, D. Bagayoko, and E. L. Kelley, Misconceptions and the certainty of response index (CRI), Phys. Educ. **34**, 294 (1999).
[9] P. Klein, S. Küchemann, S. Brückner, O. Zlatkin-Troitschanskaia, and J. Kuhn, Student understanding of graph slope and area under a curve: A replication study comparing first-year physics and economics students, Phys. Rev. Phys. Educ. Res. **15**, 020116 (2019).
[10] M. Planinic, W. J. Boone, R. Krsnik, and M. L. Beilfuss, Exploring alternative conceptions from newtonian dynamics and simple DC circuits: Links between item difficulty and item confidence, J. Res. Sci. Teach. **43**, 150 (2006).
[11] M. Potgieter, E. Malatje, E. Gaigher, and E. Venter, Confidence versus performance as an indicator of the presence of alternative conceptions and inadequate problem-solving skills in mechanics, Int. J. Sci. Educ. **32**, 1407 (2010).
[12] I. Testa, A. Colantonio, S. Galano, I. Marzoli, F. Trani, and U. Scotti Di Uccio, Effects of instruction on students' overconfidence in introductory quantum mechanics, Phys. Rev. Phys. Educ. Res. **16**, 010143 (2020).
[13] M. M. Hull, A. Jansky, and M. Hopf, Two approaches to analyzing fluidity of student reasoning on a multiple choice survey about half-life, Electronic Proceedings of the ESERA 2021 Conference (to be published).
[14] M. M. Hull and M. Hopf, Student understanding of emergent aspects of radioactivity, Int. J. Phys. Chem. Educ. **12**, 19 (2020), https://ijpce.org/index.php/IJPCE/article/view/125.
[15] M. M. Hull, A. Jansky, and M. Hopf, Reasoning fluidly about half-life on a two-tier multiple-choice survey, GDCP Conf. Proc. 2020, virtual conference.
[16] M. M. Hull, Emergent Aspects of Radioactivity: Creation of a Survey on Half-life, *Proceedings of the GDCP (Gesellschaft für Didaktik der Chemie und Physik) 2019 Conference*, pp. 590–593, https://www.gdcp-ev.de/wp-content/tb2020/TB2020_590_Hull.pdf.
[17] C. Singh, Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer, Phys. Rev. ST Educ. Res. **4**, 010105 (2008).
[18] H. M. C. Eijkelhof, Radiation and risk in physics education, *Radiation and Risk in Physics Education* (CD[beta] Press, Rijksuniversiteit Utrecht, Netherlands, 1990).

[19] P. L. Lijnse, H. M. C. Eijkelhof, C. W. J. M. Klaassen, and R. L. J. Scholte, Pupils' and mass-media ideas about radioactivity, Int. J. Sci. Educ. **12**, 67 (1990).
[20] H. Eijkelhof and R. Millar, Reading about Chernobyl: The public understanding of radiation and radioactivity, Sch. Sci. Rev. **70**, 35 (1988).
[21] E. E. Prather, Students' beliefs about the role of atoms in radioactive decay and half-life, J. Geosci. Educ. **53**, 345 (2005).
[22] M. M. Hull, A. Jansky, and M. Hopf, Probability-related naïve ideas across physics topics, Stud. Sci. Educ. **1**, 45 (2021).
[23] C. W. J. M. Klaassen, H. M. C. Eijkelhof, and P. L. Lijnse, Considering an alternative approach to teaching radioactivity, *Relating Macroscopic Phenomena to Microscopic Particles: A Central Problem in Secondary Science Education* (CD-Beta Press, Rijksuniversiteit Utrecht, Netherlands, 1990), pp. 304–316.
[24] A. Jansky, Die Rolle von Schülervorstellungen zu Wahrscheinlichkeit und Zufall im naturwissenschaftlichen Kontext, *Die Rolle von Schülervorstellungen Zu Wahrscheinlichkeit Und Zufall Im Naturwissenschaftlichen Kontext, University of Vienna*, 2019.
[25] U. Wilensky and M. Resnick, Thinking in levels: A dynamic systems approach to making sense of the world, J. Sci. Educ. Technol. **8**, 3 (1999).
[26] D. Hammer, More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research, Am. J. Phys. **64**, 1316 (1996).
[27] D. Hammer, Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions, J. Learn. Sci. **5**, 97 (1996).
[28] D. Hammer, Student resources for learning introductory physics, Am. J. Phys. **68**, S52 (2000).
[29] D. Hammer, A. Elby, R. E. Scherr, and E. F. Redish, Resources, framing, and transfer, in *Transfer of Learning from a Modern Multidisciplinary Perspective*, edited by J. P. Mestre (IAP, Charlotte, NC, 2006), pp. 89–121.
[30] S. Kapon and A. A. diSessa, Instructional explanations as an interface—the role of explanatory primitives, AIP Conf. Proc. **1289**, 189 (2010).
[31] J. Minstrell, Facets of students' knowledge and relevant instruction, *Research in Physics Learning: Theoretical Issues and Empirical Studies* (IPN, Kiel, 1992), pp. 110–128.
[32] J. P. I. Smith, A. A. diSessa, and J. Roschelle, Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition, J. Learn. Sci. **3**, 115 (1994).
[33] H. A. Simon and J. R. Hayes, The understanding process: Problem isomorphs, Cogn. Psychol. **8**, 165 (1976).
[34] J. R. Hayes and H. A. Simon, Psychological differences among problem isomorphs, in *Cognitive Theory*, edited by N. J. Castellan, D. B. Pisoni, and G. R. Potts (Lawrence Erlbaum, Hillsdale, New Jersey, 1977).
[35] K. Kotovsky, J. R. Hayes, and H. A. Simon, Why are some problems hard? Evidence from the Tower of Hanoi, Cogn. Psychol. **17**, 248 (1985).
[36] C. Singh, Assessing student expertise in introductory physics with isomorphic problems. I. Performance on nonintuitive problem pair from introductory physics, Phys. Rev. ST Phys. Educ. Res. **4**, 010104 (2008).
[37] B. D. Pulford, A. M. Colman, E. Buabang, and E. M. Krockow, The persuasive power of knowledge: Testing the confidence heuristic, J. Exp. Psychol. Gen. **147**, 1431 (2018).
[38] J. Kruger and D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, J. Pers. Soc. Psychol. **77**, 1121 (1999).
[39] B. A. Lindsey and M. L. Nagel, Do students know what they know? Exploring the accuracy of students' self-assessments, Phys. Rev. ST Phys. Educ. Res. **11**, 020103 (2015).
[40] J. Leppavirta, Assessing undergraduate students' conceptual understanding and confidence of electromagnetics, Int. J. Sci. Math. Educ. **10**, 1099 (2012).
[41] D. E. Brown and J. Clement, Overcoming misconceptions via analogical reasoning: Abstract transfer versus explanatory model construction, Instr. Sci. **18**, 237 (1989).
[42] B. Keogh, S. Naylor, and C. Wilson, Concept cartoons: A new perspective on physics education, Phys. Educ. **33**, 219 (1998).
[43] J. Woithe, Designing, measuring and modelling the impact of the hands-on particle physics learning laboratory s'cool lab at cern effects of student and laboratory characteristics on high-school students' cognitive and affective outcomes, Ph.D. thesis, Kaiserslautern University, 2020.
[44] R. Millar, School students' understanding of key ideas about radioactivity and ionizing radiation, Publ. Understand. Sci. **3**, 53 (1994).
[45] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.020108 for the full SWORDE survey as well as details on its creation and validation.
[46] M. M. Hull, A. Jansky, and M. Hopf, Radioactivity as "quintessentially eternal": two survey prompts, *Proceedings of the GDCP (Gesellschaft für Didaktik der Chemie und Physik) 2021 Conference*, pp. 184–187, https://gdcp-ev.de/wp-content/uploads/2022/05/Tagungsband-2022-Stand-13522.pdf.
[47] P. Mayring, Qualitative content analysis: Theoretical foundation, basic procedures and software solution, *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution* (Klagenfurt, 2014).
[48] R. L. Thorndike, Who belongs in the family?, Psychometrika **18**, 267 (1953).
[49] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).
[50] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).
[51] B. R. Wilcox and H. J. Lewandowski, Developing skills versus reinforcing concepts in physics labs: Insight from a survey of students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **13**, 010108 (2017).
[52] S. Jamieson, Likert scales: How to (ab)use them, Med. Educ. **38**, 1217 (2004).

[53] G. Norman, Likert scales, levels of measurement and the "laws" of statistics, Adv. Health Sci. Educ. **15**, 625 (2010).

[54] G. Zhang, T. J. Anderson, M. W. Ohland, and B. R. Thorndyke, Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study, J. Eng. Educ. **93**, 313 (2004).

[55] J. P. Zwolak, R. Dou, E. A. Williams, and E. Brewe, Students' network integration as a predictor of persistence in introductory physics courses, Phys. Rev. Phys. Educ. Res. **13**, 010113 (2017).

[56] M. D. F. Fuad, B. M. N. Al-Zurfi, M. A. AbdulQader, M. F. A. Bakar, M. Elnajeh, and M. R. Abdullah, Prevalence and risk factors of stress, anxiety and depression among medical students of a private medical University in Malaysia, Educ. Med. J. **7**, e1 (2015).

[57] E. A. Dare and G. H. Roehrig, "If I had to do it, then I would": Understanding early middle school students' perceptions of physics and physics-related careers by gender, Phys. Rev. Phys. Educ. Res. **12**, 020117 (2016).

[58] D. McPadden and E. Brewe, Impact of the second semester university modeling instruction course on students' representation choices, Phys. Rev. Phys. Educ. Res. **13**, 020129 (2017).

[59] J. Cohen, Statistical power analysis for the behavioral sciences, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).

[60] E. F. Redish, Teaching physics with the physics suite, *Teaching Physics with the Physics Suite* (Wiley, New York, NY, 2003).

[61] I. S. Caleon and R. Subramaniam, Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions, Res. Sci. Educ. **40**, 313 (2010).

[62] A. Maries, N. I. Karim, and C. Singh, Active learning in an inequitable learning environment can increase the gender performance gap: The negative impact of stereotype threat, Phys. Teach. **58**, 430 (2020).