Exploring student estimates of astronomical scales: Impact of question formulation and visualization

Willem Keppens,[*] Mieke De Cock, and Hans Van Winckel
Department of Physics and Astronomy, KU Leuven, Celestijnenlaan 200C, 3001 Leuven, Belgium

Wim Van Dooren
Faculty of Psychology and Educational Sciences, KU Leuven, Dekenstraat 2, 3000 Leuven, Belgium

Jan Sermeus
Department of Physics and Astronomy, KU Leuven, Celestijnenlaan 200C, 3001 Leuven, Belgium;
Faculty of Psychology and Educational Sciences, KU Leuven, Dekenstraat 2, 3000 Leuven, Belgium;
and Royal Observatory of Belgium, Planetarium, Bouchoutlaan 10, 1020 Brussels, Belgium

(Received 19 February 2025; accepted 19 May 2025; published 11 June 2025)

Estimating astronomical scales requires multiple complex mental processes, such as spatial thinking and interpreting large numbers. As such, it is a nontrivial question how these estimates can be most efficiently assessed. There is reason to believe that results from previous studies probing astronomical scale estimates are possibly susceptible to effects inherent in the questioning methods used. In this study, an interactive online survey was constructed and administered to 201 students in their last year of high school. To probe their estimates of spatial scales in the Solar neighborhood, we formulated five questions, two of those probing estimates on the relative sizes of astronomical bodies and three on the relative distances between those bodies. These questions were formulated in two different ways, and the effect of these formulations was studied. In one formulation, students were asked to numerically compare the magnitude of two sizes or distances, while in the other, these estimates were made in terms of the travel time of an imaginary spacecraft. After every answer, students were confronted with a customized visualization, which they could either agree with and move on to the next question or disagree and reconsider their previous answer until they agreed. Studying the effect of these visualizations on the students' answers was another objective of this work. Most students had difficulties estimating both the relative sizes of the considered celestial bodies and the distances between them. The range of estimates covered many orders of magnitude for all questions, and for the distance-related questions, there was a clear trend of underestimation. We found a significant impact of question formulation on the magnitudes of student estimates. However, we found no indication that one question formulation led to more reliable results than the other. The effect of the visualizations was smaller than anticipated but noticeably larger for the size-related questions than for the distance-related questions. Self-assessments of certainty were made by the students after every answer, and those were found not to correlate with the accuracy of the answers. The implications of these findings are discussed.

DOI: 10.1103/cv8n-mshc

I. INTRODUCTION

In our everyday lives, we often benefit from accurate estimates of sizes and distances. By experience, we become more skilled in interpreting the length of a 80 km trip or the width of a 21 cm pizza. However, when the involved scales extend beyond the boundaries of Earth, people can not rely on personal experience anymore to make such estimates. This is only one reason why it is so difficult for most (if not all) people to make estimates on astronomical scales [1]. Another reason is the absolute vastness of the universe: how to express an estimate for a distance you only know to be "very, very large," without any upper limit in mind? If these scales are to be expressed in familiar units like kilometers, then quickly we risk dealing with unfamiliarly large numbers, and it has been shown that students have difficulties grasping such large numbers [2–5]. Alternatively, one could turn to alternative units more regularly used in astronomical contexts, like lightyears,

[*] Contact author: willem.keppens@kuleuven.be

but also those concepts have been shown to confuse students [6,7]. Nonetheless, having a realistic idea of astronomical scales is important for several reasons. Not only does it enhance one's appreciation for the vastness of space and the relative smallness of Earth, but it also stimulates an accurate interpretation of astronomical phenomena [8,9] and plausibly helps put astronomical images and discoveries into a more correct perspective. For instance, when new exoplanets are being discovered in the "vicinity" of the Solar System, a realistic comprehension of interstellar distances enhances the understanding that those planets are not actually next door. Moreover, reasoning about the vast scales involved in astronomy requires spatial thinking skills [10], which are of great use in multiple science, technology, engineering, and mathematics disciplines and are strongly correlated with logical thinking [11]. Investigating students' estimates on astronomical scales has therefore not only relevance in Astronomy Education Research but also potentially in the adjacent field of Physics Education Research. However, despite its importance, astronomical scales are hardly covered in any high school classroom. As Lelliott and Rollnick [8] write in their review on astronomy education research over the period between 1974 and 2008: "Finally, there should be a greater focus on the teaching of distance and size to help explain astronomical phenomena. Although very few studies focused on this big idea, it is crucial to so much of astronomy, from the size of the Earth and the Solar System to their relationship to the rest of the Galaxy and the Universe. Not only is this concept under-researched, but it is under-taught."

The above illustrates the importance of scale comprehension and why it is very difficult for most people to make estimates on astronomical scales. Additionally, assessing such estimates unambiguously is also a nontrivial task from a research point of view. This becomes clear from previous studies with children, adolescents, and adults, e.g., [12–18]. Those studies varied in the precise questioning methods used, and while some findings are common, others differ significantly between sources. In some cases, there is reason to believe that the results of the surveys were influenced by the precise formulation of the questions. We will elaborate more on this point in Sec. II A 3, with concrete examples.

This study has three main objectives: First, we aim to probe student estimates of spatial scales in the Solar neighborhood by means of an interactive online survey. These estimates are thought to provide insight into students' underlying mental models of the Solar System. Second, we investigate the effect of different question formulations on students' answers. Third, we examine to what extent these estimates are altered when the students are confronted with a visualization of their answers.

We discuss current insights from the literature on the astronomical knowledge of students and their perception of astronomical scales in Sec. II. Moreover, we also specify what is meant by the notion of mental models, and why they can be expected to have a visual character. Then, in Sec. III, we describe how this study was set up. In Sec. IV, we summarize the main results of this work, and in Sec. V, we provide an extensive discussion of those results.

II. BACKGROUND

A. Astronomical knowledge of students

1. Misconceptions in astronomy

When students enter a classroom, their minds are not simply blank slates upon which new, scientific information can be written. This was first mentioned by Piaget, who discussed the notion of "preconcepts" to address children's intuitive understanding of the world, which is based on their everyday experience [19]. Since then, several other terms have been used to describe these student ideas, such as alternative concept(ion)s [20,21], alternative frameworks [22], intuitive theories [23], and misconceptions [24,25] (the latter being the term we will use throughout this text). Despite subtle differences in their exact meaning, there is a general consensus that "this intuitive knowledge provides explanations of natural phenomena which are frequently different from the current accepted scientific explanations and which tend to be resistant to change" [26].

The field of astronomy is no exception to this, as numerous astronomical misconceptions have by now been identified and described. Many of these misconceptions are not limited to one specific geographical region or demographic group and can be held by children, adolescents, and adults alike. A few examples of common misconceptions are the Moon's phases are caused by the Moon moving into the Earth's shadow [13,27–30], the Sun always rises exactly at due east [29–32], stars are fixed and unmoving as seen from Earth [33,34], the Sun is at the center of the Universe [13,27,35,36], the day-night cycle is caused by the revolution of the Sun around the Earth [13,28]. However, of more direct relevance to this work are astronomical misconceptions related to scales, which are discussed in the next section.

2. Misconceptions related to scale

Many astronomical misconceptions are closely related to the sizes of celestial bodies and the distances between them. For example, people tend to believe that the stars are closer to Earth than Pluto, as Trumper showed with both university students [37] and preservice teachers [13]. Lightman and Sadler [38] obtained similar results, as less than half of their participating students could rank the terms "space shuttle in orbit," "planets," and "stars" correctly in increasing distance from Earth. The authors conclude that students estimate astronomical distances based on visual clues, placing brighter stars closer than dimmer planets, without realizing that the former may be larger and brighter but farther away.

Several authors have reported flawed estimates on astronomical scales made by their participants. For example, many students believe that the Sun is the largest and/or hottest star in our Universe (as shown with pupils between 12 and 18 years old by, e.g., Serttaş and Türkoğlu [36] and Bitzenbauer *et al.* [39]). Furthermore, students systematically underestimate the vast distances in space, such as the distance between the Earth and the Moon [14], the Earth and the Sun [12,14,17], and the Sun and a close star [12,13,17]. While these and other sources may agree on these findings, there are also interesting differences to be noted in their results. We will elaborate more on these in Sec. II A 3. The above examples illustrate how many students' astronomical mental models are skewed on multiple fronts.

As mentioned in the introduction, the main reason why students struggle with the estimation of astronomical scales is that they are nowhere near their everyday life experiences. It is challenging to conceptualize the vastness of our Solar System, let alone the entire Universe. When forced to make estimates on such scales, people rely on information they gathered through encyclopedia, textbooks, television, online media, or instruction at school. However, such information is almost always accompanied by visual representations that are inaccurate in terms of scale. After all, it is simply impossible to depict the whole Solar System—or even only the Sun-Earth-Moon system —to scale on a regular-sized piece of paper or digital screen in a useful way. Instead, astronomical images aim to explain a certain phenomenon such as the Moon phases by making a schematic depiction, where the main features of interest are strongly exaggerated and distances are usually minimized and disproportional. This practical deformation is of course justified, necessary even, to explain astronomical concepts and should in principle not be problematic as long as students are aware of its schematic nature. Unfortunately, that is not always the case. Testa *et al.* [40] showed that students often have difficulties interpreting astronomical textbook images, taking certain features too literally. As such, textbook images could reinforce or even create common misconceptions rather than correcting them.

For instance, one feature that is often drastically exaggerated in textbook images is the eccentricity of Earth's orbit around the Sun. Rather than highly elliptical, Earth's orbit is actually (almost) indistinguishable from a perfect circle. However, this skewed representation may mislead students [41] toward what could well be the most widely held astronomical misconception of all: the belief that the seasons are caused by a varying distance between the Earth and the Sun [8,13,27–31]. Although the link between textbook images and this particular misconception appears plausible, it should be noted that no explicit confirmation has been found in the literature. Nonetheless, the misconception is clearly triggered by an erroneous idea of an astronomical distance and the variation therein over the course of a year. The astronomical phenomenon of the seasons is not by itself related to scale, but this example neatly illustrates how skewed ideas of sizes and distances can be a cause for other, non-scale-related misconceptions. Thus, this example stresses the importance of ensuring that students develop an accurate understanding of astronomical scales.

3. Difficulties with question types

Many researchers have assessed student estimates on astronomical scales while using very different methods. For example, a number of studies used multiple choice questions to probe student estimates [12–16]. This method has the advantages that it is time efficient and that results are easily and reliably processed and analyzed. However, this format forces students to choose between the provided options, thereby reducing the flexibility to express their own estimates. Moreover, the results obtained through multiple choice questions are possibly influenced by the provided set of options. This is especially true when those options deprive participants of the possibility of under- or overestimate. For example, Sadler [12] included multiple choice questions to probe the distances between the Sun and Earth and the Sun and a close star. These questions were later copied by both Shore and Kilburn [14] and by Trumper [13] in their surveys with students and teachers in the U.S. and Israel, respectively. All these studies found varying but nonetheless significant percentages of participants underestimating both distances. However, in both questions, the correct answer was the largest option, making overestimations *a priori* impossible. When Mant repeated one of these questions while including four additional, larger options, he concluded that 50% of the participants actually overestimated the Sun-Earth distance [15]. Miller and Brewer [17] suggest that students often do not know the values of these distances but only know that they are surprisingly large. As a consequence, they choose the largest available multiple choice option, whatever that might be. This, of course, is equally undesirable as offering the correct answer as the largest response alternative.

As an alternative to the multiple choice format, open-ended questions have also been used to probe students' estimates on astronomical scales, either in a written format or with interviews (see, e.g., [42,43]). This approach provides more flexibility for students to freely express their thoughts, without imposing a choice between options. However, Miller and Brewer [17] notice that it is better to avoid students estimating astronomical scales in terms of everyday units like kilometers or miles because "this method confounds the students' intuitive conceptions about astronomical distances with their ability to understand large numbers. Furthermore, some of the students could have memorized the distance from the Earth to the Moon without having a real understanding of the meaning of the memorized information" [17]. Indeed, as already mentioned, children and adolescents are known to have difficulties grasping the meaning of large numbers. This has been shown through research on number line estimation (NLE) tasks, in which young individuals tend to indicate the position of numbers logarithmically instead of linearly (see e.g., studies by Laski and Siegler [44] and Rips [45], and references therein). When growing older, children undergo a "representational shift" from logarithmic to linear patterns of numerical estimates. This was first shown to occur in children between second and sixth grade by Siegler and Opfer [46], although not all authors agree on the theoretical explanations of these results [47,48]. Nonetheless, it is clear from these findings that any method involving large numbers to assess student estimates is far from optimal since the validity of the results is undermined by the inability of students to interpret those numbers. Also, Rajpaul *et al.* [18] conclude that it is difficult to compare the results of MC questionnaires with open-ended question studies since those results are strongly linked to how the questions are asked.

Questions in an open-ended format do not necessarily require answers with large numbers. One method to avoid these answers is to refer to a small-scale model of the celestial bodies of interest. For instance, in their research with 83 U.S. undergraduate students, Miller and Brewer [17] used an open-ended format in which they (hypothetically) represented the Earth as a baseball at the doorstep of the lecture auditorium. Students were asked to estimate the distance to several celestial bodies on this scale, expressing their answers either in some unit of distance or in terms of a specific location or landmark. In this way, they aimed to avoid the concerns of both multiple choice questions and large numbers discussed above. While this method certainly has potential, the data obtained in this way were not always very specific. When students estimated the location of objects to be "in Canada" or "out of town" for example, it was up to the authors to translate this into numbers, thereby requiring interpretation by the researchers. Moreover, this approach assumes that students have an accurate idea of distances in their direct environment, for example, those to the next town or to a neighboring country. This assumption may not be justified for all participating students.

In this work, we aim to assess student estimates of astronomical scales, while keeping in mind all the caveats discussed before. We want to investigate specifically how these estimates are influenced by two factors: the use of different question formulations and the presentation of customized visualizations. The effect of different question formulations on student answers has already been explored by Favia *et al.* [49]. They conducted a survey of astronomical statements based on common misconceptions with over 600 undergraduate students at the University of Maine. They found that the number of students proven to hold a certain misconception is significantly dependent on the precise formulation of that misconception. In this work, we investigated whether similar effects could be found while probing students' estimates of astronomical scales. We argue that these estimates are indicative of the students' mental models of the Solar System. We will elaborate more on our methodology in Sec. III. However, we first specify formally what is meant by a mental model in Sec. II B 1. In Sec. II B 2, we argue why mental models can be expected to have a visual character and how we interpret them in the context of astronomical scales.

B. Mental representations

1. Mental models

The precise definition of a mental model differs between sources. The term was first introduced in 1943 by psychologist Kenneth Craik, who stated that "people carry in their minds a small-scale model of how the world works" [50]. Since then, many authors have suggested their own descriptions and interpretations of what a mental model entails precisely. For example, Johnson-Laird [51] writes that "mental models are structural analogs of the world as perceived and conceptualized." In the context of system dynamics, Doyle and Ford [52] state that "A mental model […] is a relatively enduring and accessible but limited internal conceptual representation of an external system whose structure maintains the perceived structure of that system." In the same article, Doyle and Ford give an extensive review of various definitions used in the literature. Ubben [53] argues that "mental models are individual types of mental modal patterns that possess a functional potential and are based on outside experiences." In a more metaphorical sense, van Ments and Treur [54] describe mental models as "a kind of blueprints or pictures in the mind that can occur in various forms."

In essence, mental models refer to internal representations that people form of the outside world through their interaction with it [32]. In this article, we refer to these internal representations whenever we make mention of mental models, without losing ourselves in the semantics of the exact terminology. Furthermore, we follow the ideas of Corpuz and Rebello [55], which suggest that mental models are private in nature and of Gilbert and Boulter [56], which state that they are essentially inaccessible to researchers. Therefore, we can only obtain hints of its shape and contents through a concrete expression of it, to which the previous authors refer as "expressed models" [55,56].

Expressed models can provide valuable insights into the underlying knowledge structure from which they are generated [57]. Despite numerous studies on this topic, no real consensus about the nature of this knowledge structure has been reached. Instead, two prominent but competing perspectives coexist regarding knowledge structure coherence: knowledge-as-theory perspectives and knowledge-as-elements perspectives [58]. Proponents of the former argue that intuitive knowledge consists of a coherent and systematic set of ideas, which can therefore most accurately be described as an intuitive theory [26]. Advocates of the latter, on the other hand, state that cognitive structures are made up of small individual units of knowledge, which are incoherent in nature and often cued by contexts [59]. These units of knowledge were named phenomenological primitives, or p prims for short [60]. Although these two views on internal knowledge structures have long appeared incompatible, recent developments reveal progress toward reconciling the two theories. For example, Vosniadou and colleagues [61] suggested that the framework theory approach can be less cohesive than initially suggested and allow for the presence of p prims in our knowledge system.

2. Visual representations

There are several indications that mental representations, especially when connected to spatial scales, have a visual aspect. For example, students' understanding of magnification and scale is tightly linked to spatial visualization [11]. The study of Konkle and Olivia [62] gave people memory, imagery, and perceptual preference tasks to determine whether objects of different physical sizes also had a different preferred visual size across participants. They found that real-world objects indeed have a so-called canonical visual size at which they are preferentially imagined, drawn, and viewed. Moreover, it is stated that this preferred imaginary size is proportional to the logarithm of the assumed actual size of the object in the world.

Mental scaling and mental rotation involve cognitive simulations of changing the size or orientation of a representation of spatial stimuli [63,64]. Support for this claim has been provided by several studies. For example, multiple authors have shown how the time it takes to imagine a rotated object is proportional to the degree of rotation of that object (see, e.g., [65,66]). Similarly, the time required to estimate the size of a magnified object is proportional to the degree of magnification. This was shown in a study by Szubielska and Balaj [63], who made participants explore a toy object either visually or tactilely and subsequently estimate the size of that object after a certain scaling was applied. This estimate had to be expressed either verbally or bimanually. The time necessary to make a size estimate was found to increase with the degree of rescaling, not influenced by the modality of perception. They also showed that the estimation accuracy became significantly lower with decreasing scale, but only when the estimate was made bimanually and not verbally. These results support the pictorial view on mental scaling, which is most commonly explained through the metaphor of a magnifying glass used to "zoom" into or out of an object.

Eye tracking studies showed how the cognitive experience of making a mental representation shares many similarities with the experience of visual perception [67]. Johansson *et al.* [68] tracked the eye movements of participants while listening to or retelling a story or describing a picture, both in front of a whiteboard and in complete darkness. They conclude that eye movements reflect the positions of objects in all these tests. The uncovered effect was equally apparent for visual and auditory stimuli, and occurred both when watching a whiteboard and while staring into absolute darkness. In a similar study, French participants were asked to imagine a map of France and name as many cities or towns as they could. Again, eye movements correspond to the physical location of the mentioned cities [69]. These studies provide evidence for the analogical theory of mental imagery, which states that "mental representations have a pictorial nature that preserves the spatial characteristics of the environment that is mentally represented" [67].

These indications for the visual (pictorial) nature of a mental representation stimulate a careful evaluation of any questioning method probing astronomical scales. The studies discussed above suggest that when students are asked to estimate a certain astronomical size or distance, they are likely to construct a visual mental image upon which they base their answers. For example, when asked to compare the sizes of the Sun and the Earth, the student is thought to construct a visual mental image of both objects with what he believes to be the correct magnitude ratio. Of course, this mental image can be scientifically correct but also completely wrong. However, the translation of this mental image into a numerical value—in this example, into the ratio of the Sun's radius to the Earth's radius—could very well induce a second source for error, in addition to the error on the image itself. We argue that this second error source is undesirable for research purposes, as it only "blurs" the expressed mental model that we aim to study. Therefore, our study included customized visualizations for all questioned relative sizes and distances, to which the students could compare their mental images. In this way, we hope to make the students' answers match their mental image more closely.

III. METHODOLOGY

A. Online survey

The goal of the online survey was to capture students' mental models about sizes and distances in space, and the influence of different question formulations and visualizations on these models. From these influences (if any), we aimed to extract information on the stability of the mental models. As both the available time in classrooms and the concentration timespan of students are always limited, a concise set of relevant sizes and distances had to be selected. To prevent extending our questionnaire toward too extreme astronomical sizes and distances, we opted to focus on the Solar System and its celestial bodies. An exception was made for the distance to Proxima Centauri, the “nearest next star” to the Sun. The probed distances were therefore:

1. Earth-Moon.
2. Sun-Earth.
3. Sun-Neptune.
4. Sun-nearest next star.

The choice for the first two distances is self-evident since the Moon and the Sun are the two most apparent bodies in the sky. The third distance was chosen as a proxy for the size of our Solar System, and the last distance was intended as a general indication of the distances between different stars.

The sizes of interest were very much aligned with the above distances. They were focused on the diameter of the following objects:

1. Moon.
2. Earth.
3. Sun.
4. Nearest next star.

Obviously, estimating the size of Proxima Centauri is a very difficult question for students. The goal of including it here was not to obtain estimates on its size specifically, but rather to check whether students realize that although stars look small in the sky, they are actually all large bodies more comparable to the size of the Sun than to that of the Moon, Earth or other planets. We elaborate more on how we approached this in Sec. III A 1.

Instead of questioning the absolute magnitudes of these sizes and distances, we opted to assess their magnitudes relative to each other. The motivation for this choice is twofold. First, according to Tretter *et al.* [70], conceptions of relative scale are often more accurate than those of absolute scale. Their study investigated the size estimates of 215 participants, ranging from fifth grade to doctoral students. After having the participants estimate the size of various objects in both absolute and relative terms, the authors were able to deduce that information about relative size was more readily understood than exact size. Second, by questioning relative sizes and distances, we avoid the difficulties of very large numbers that we addressed in Sec. II A 3. We thus arrive at five ratios that we ask students to estimate, three of which relate to distances and two relate to sizes. They are:

1. $\frac{\text{Sun-Earth}}{\text{Earth-Moon}}$.
2. $\frac{\text{Sun-Neptune}}{\text{Sun-Earth}}$.
3. $\frac{\text{Sun-nearest next star}}{\text{Sun-Neptune}}$.
4. $\frac{\text{Earth}}{\text{Moon}}$.
5. $\frac{\text{Sun}}{\text{Earth}}$.

These ratios are approximately 400, 30, 9000, 4, and 109, respectively. Note that these are all relatively small numbers, presumably well comprehensible for the participants in this study.

A third consideration was whether the students had to estimate the magnitude of the larger size/distance compared to the smaller one or vice versa. Studies on number line estimations (NLE) have pointed out that these two tasks require different mental processes and solution strategies [71]. In bounded NLE tasks, the left and right edges of a line are both indicated with a number, and participants have to indicate the position of a third number in between (e.g., marking 37 on a 0–100 number line). In the unbounded NLE tasks, only the left edge has an indicative number, and a second number is drawn closely next to it as a unit measure of magnitude. Participants then indicate the position of a third number, which is larger than both indicated numbers (e.g., marking 37 on a number line that has marks for 0 and 1). Bounded NLE is based on proportion judgment, while unbounded NLE is magnitude-estimation-based [71]. Since the sizes and distances of interest in this study are presented as relative proportions, and since in an earlier pilot study a sample of students—comparable to the participants of the main study—performed quite well in a bounded NLE task,[1] we opted to work from larger sizes or distances toward smaller ones. What we mean by this is that in the first quantification question, participants started from the largest ranked size or distance and were asked to comparatively estimate the magnitude of the second largest ranked size or distance. Thereafter, the second largest size or distance was used as a measure to which the magnitude of the third largest size or distance had to be estimated, and so on. As such, participants were always asked to estimate the magnitude of a smaller size or distance (the denominators in the above listing) compared to a larger one (the numerators).

1. Structure and question formulations

Prior to estimating the magnitudes of the five ratios of interest, students were asked to rank the relevant sizes and distances from smallest to largest. Only thereafter, the ranked sizes and distances were evaluated pairwise and the ratios were quantified. This quantification was done according to the ranking answers of every student; if a student incorrectly ranked the Sun-Neptune distance as smaller than the Sun-Earth distance, then the subsequent quantification question would probe how much smaller the former is than the latter.

For the considered sizes, the nearest next star was included in the initial ranking questions of the survey but omitted for the further quantification questions. The goal was to see if the student realized that since it is a star, it must surely be ranked larger than both the Moon and the Earth. This is not a reasoning that can be assumed to be shared by all participants, as the literature shows that many students believe that stars are very small due to their tiny appearance in the night sky [43,72]. Whether the star was

[1] We elaborate more on the performed pilot studies in Appendix A.

ranked smaller or larger than the Sun was not of primary importance.

The distance "Sun-nearest next star" was included both in the initial ranking question for distances and in the subsequent quantification questions. This inclusion was done because we want to probe for the misconception that the next star lies within our Solar System [7,13,27], but we also want to assess the distance between stars compared to the size of a planetary system (here: the distance Sun-Neptune) for students who do not hold that misconception.

We asked all students to estimate the relevant ratios twice, using two different question formulations. The first formulation, which we will henceforth name the "fraction" formulation, is a straightforward numerical comparison of the two sizes or distances. An example of a question in this formulation is: "How many times smaller is the distance Earth-Moon compared to the distance Sun-Earth?" To avoid confusion, an example was added to these questions. For the distance questions, that would be: "Example to clarify: the distance Brussels-Antwerp (±50 km) is approximately 4 times smaller than the distance Brussels-Amsterdam (±200 km)." For the size questions, the fraction formulation would refer to the radii of the two objects to compare. An example of such a question is: "How many times smaller is the radius of the Earth compared to the radius of the Sun? Example to clarify: the radius of a tennis ball (±3 cm) is about 4 times smaller than the radius of a basketball (±12 cm)." By assessing estimates on relative sizes in terms of radius and not in terms of area or volume, we restrict the magnitude of the estimates as much as possible, thus avoiding methodological caveats concerning large number interpretation.

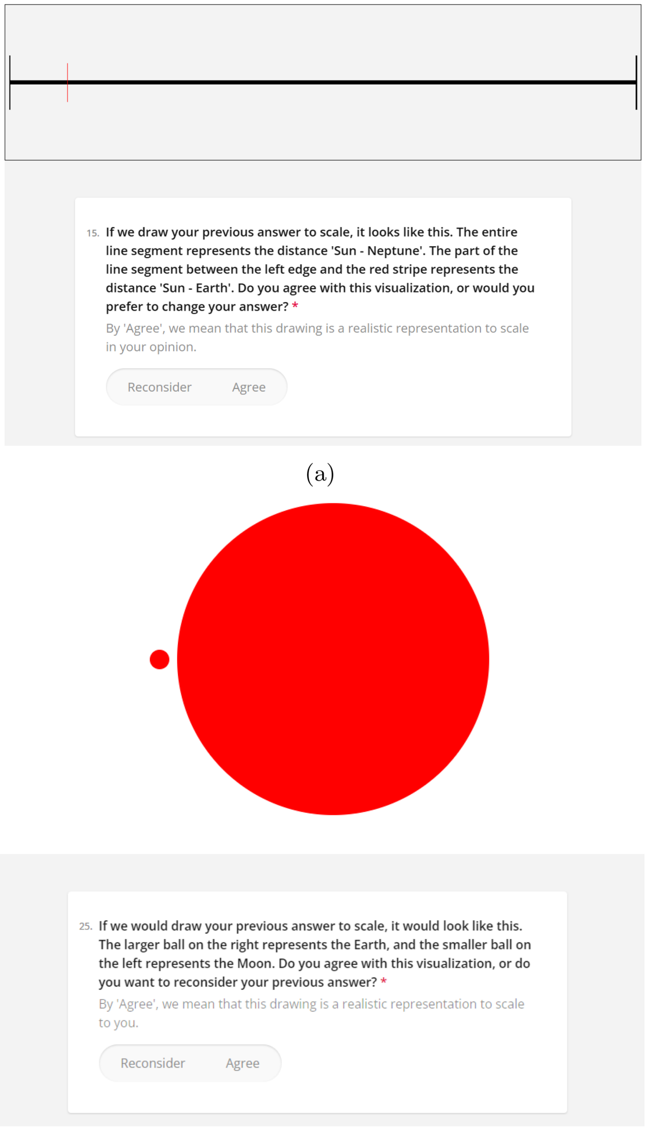
As a second question formulation, the same ratios were probed with a formulation about the travel time of an imaginary space rocket. We will refer to this formulation as the "travel time" formulation. The previous distance question would read in this formulation: "Imagine that it takes an imaginary rocket one week to travel the distance Sun-Earth. Then how long would it take this rocket to travel the distance Earth-Moon?" Students were free to choose in which unit(s) of time they wanted to answer, as long as the total time of the answer was smaller than the travel time to compare with (in this example, smaller than one week). For size questions, students were asked to estimate the travel time necessary to circumnavigate the celestial bodies. Although this may not appear to be an obvious or intuitive formulation to assess these sizes, it was chosen because it felt more natural than the travel times necessary to cover the bodies' radii, while still preserving a close analogy to the distance questions. Half of the participating students first answered the questions in the fraction formulation and then in the travel time formulation, while the other half worked in the opposite order.

The mental processes to answer the two question formulations can be expected to differ drastically. The fraction formulation allows one to express a relative magnitude as a dimensionless number, and no calculations are required. However, with this formulation, we are not entirely exempted from the difficulties concerning large number interpretation. Although we are not working with distance units like kilometers, the abstractness of the large magnitudes of estimates may still be problematic for students. This is especially true for the ratio $\frac{\text{Sun-next star}}{\text{Sun-Neptune}}$, with a scientific value of 9000.

The advantage of the travel time formulation is that this abstractness is completely erased. The travel times involved are very concrete, ranging from roughly a second (or fractions of a second, if students underestimate) to a few years. It could be argued that students are likely to have a better intuitive understanding of how small one minute is compared to a week than of the corresponding fraction, roughly $1/10000$. This hypothesis is supported by the results of Makwela [73], who found that students tend to spontaneously express large distances in terms of a journey and (travel) time. Another feature of the travel time formulation is that it is difficult to make calculations in the minutes-hours-weeks system. Even if a student knows that the ratio $\frac{\text{Sun-Earth}}{\text{Earth-Moon}}$ equals roughly 400, he can still run into trouble when estimating a travel time for Earth-Moon compared to a travel time of one week for Sun-Earth. However, this should not be seen as a mere disadvantage, since performing such calculations is not an objective of the survey itself. We want to obtain expressions of the mental models of the students, not to measure their calculation skills. In the example above, the student can demonstrate his mental model by first estimating an appropriate travel time and thereafter, if needed, by adjusting the subsequent visualization until it displays an appropriate ratio.

2. *Customized visualizations*

After each response, a customized visualization was shown to the student. This visualization was automatically generated by the JavaScript software running the survey. For the distance-related questions, this visualization consisted of a simple line segment representing the larger distance, and a supplementary red line indicating the smaller distance as compared to the larger one. For the size-related questions, the visualization instead showed two disks, representing the relative size of the two celestial bodies. An example of these visualizations is shown in Fig. 1, both for distances [Fig. 1(a)] and for sizes [Fig. 1(b)]. When confronted with this concrete to-scale model of their answer, students could either agree with it and move on to the next question or disagree and reconsider their previous answer. As was explained in the survey, to "Agree" was to indicate that, according to the student, the shown visualization was a realistic representation of the astronomical situation as they imagined it. If students disagreed, they could change the previous answer as many times as necessary and see the visualization change in real



15. If we draw your previous answer to scale, it looks like this. The entire line segment represents the distance 'Sun - Neptune'. The part of the line segment between the left edge and the red stripe represents the distance 'Sun - Earth'. Do you agree with this visualization, or would you prefer to change your answer? *

By 'Agree', we mean that this drawing is a realistic representation to scale in your opinion.

(a)

25. If we would draw your previous answer to scale, it would look like this. The larger ball on the right represents the Earth, and the smaller ball on the left represents the Moon. Do you agree with this visualization, or do you want to reconsider your previous answer? *

By 'Agree', we mean that this drawing is a realistic representation to scale to you.

(b)

FIG. 1. Illustrations of the visualizations as shown in the online survey for (a) distance-related questions and (b) size-related questions, translated to English.

time. However, students must eventually always click “Agree” before being able to continue to the next question. For every question, both the initial answer (before any visualization was shown) and the final answer corresponding to the agreed visualization were stored in the database.

The goal of these visualizations was to distinguish between what students initially answered and what they actually meant, that is, what their mental representations actually looked like. As discussed in Sec. II B 2, mental representations have been shown to be (at least partly) visual in nature. Therefore, matching their mental image to this concrete visual image may be more natural for students than matching it to a written answer only. Moreover, providing this visualization may partly resolve the difficulties inherent to the two question formulations. Even if students have trouble with the abstractness of the fraction formulation or with the nondecimal metric of the travel time formulation, seeing and interacting with the visualization may still guide their initial answers toward their true mental representation. Therefore, if we would find a difference in our results between the two question formulations, we could expect this difference to be larger in the initial answers than in the answers after visualization.

3. Certainty assessments

After (eventually) agreeing with the presented visualization, students were asked to self-assess their certainty. This was done on a (quasi)continuous sliding bar, where the far left corresponded to “very unsure” and the far right to “very sure.” Alternatively, students could also choose a “pure guess” option, in which case they were not asked to give any indication of certainty on the sliding bar. This option was included to distinguish purely random answers from very uncertain answers, the latter of which could still be based on some reasoning. For example, a student remembering from class that the Sun is much farther from Earth than the Moon could drastically overestimate the ratio $\frac{\text{Sun-Earth}}{\text{Earth-Moon}}$ and be very uncertain of it. However, this answer would not be a completely random guess because this student knows that e.g., a factor of 2 will surely not be a realistic estimate.

Since students could always be quite sure that their estimate would not be exactly correct, the phrasing of the certainty assessment question had to be somewhat nuanced. We did this using the following phrasing: “How certain are you that the scientifically correct answer to this question lies between [A] and [B]?” In all but one question, the software automatically calculated the values as 50% and 150% of the student answer, respectively. For example, when students estimate the ratio $\frac{\text{Sun-Earth}}{\text{Earth-Moon}}$ at a value of 100, they are then asked how certain they are that the correct answer lies somewhere between 50 and 150. Only for the question probing the relative distance between the Sun and the next star and the Sun and Neptune, the calculation of A and B was altered to 10% and 1000% of the student's answer, because this question was anticipated to be more difficult than the others, and the correct value was by far the largest of all assessed ratios.

B. Sample

The sample for this study consisted of 201 students in their last year of high school. The students were approached by their teachers, and no specific selection criteria for participation were applied. There were 64 boys and 133 girls among the participants, and 4 students indicated either to be nonbinary or that they would rather not say. Most students were between 16 and 18 years old ($\mu = 17.2$; $\sigma = 1.0$). The data were gathered in seven different schools in three provinces across Flanders, Belgium. The participating classes were informed on the subject of the survey at the beginning of the lecture. They were informed that participation was completely voluntary, that no rewards would be granted upon completion of the test, and that their answers would in no way impact their grades. A written consent was signed by both the student and the researcher, and verbal clarification on its content was provided. The students could all enter the online survey through a URL link and complete the task on their own (school) laptops. They could take as long as they needed to complete the survey, which in practice took between 20 and 30 min.

C. Analysis method

The survey was constructed using JavaScript code, written in a Visual Studio Code editor (version 1.86.1). When finished, the survey was deployed to an online server on Netlify, and connected to an online database using MongoDB Atlas, version 7.0.12. As such, a whole class of students could fill in the survey simultaneously, and when they were finished their answers were stored immediately in the online database.

From the MongoDB platform, the data were copied into a Notepad++ file and from there imported into a Jupyter Notebook worksheet. All the analysis presented in this paper was performed in that environment, coding in Python version 3.11.5. Before the quantitative analysis could start, the strings were cleaned from unnecessary symbols like "{}," and students who had skipped certain questions—because they immediately agreed with a given visualization—were assigned "None" values for those questions.

Descriptive analyses were performed to gain insight into the characteristics of the data. The distributions of the estimates to each question were described in terms of medians and interquartile ranges rather than with means and standard deviations, to neutralize the effect of extreme outliers. To compare two sets of data, we used statistical tests to decide on the significance of uncovered differences. Depending on the precise datasets to be compared, we used a binomial test, a t test, a Mann-Whitney U test, or a Wilcoxon test. The choice of the statistical test used will be mentioned and motivated with any specific result. Statistical significance is always assessed depending on p values. We differentiate between $p > 0.05$ (insignificant), $p < 0.05$ (weakly significant), $p < 0.01$ (moderately significant), and $p < 0.001$ (strongly significant).

We present the results of the ranking questions in Sec. IV A 1. We both investigate the correctness of the size (distance) ranking as a whole and of every size (distance) element individually. The answers to the quantification questions are displayed the first time in Sec. IV A 2. To discuss the effect of the customized visualizations, the sets of answers before and after the visualizations are compared in Sec. IV B, both in terms of the magnitude of the answers and in terms of their accuracy. Due to the coupled nature of the data, certain results differentiating between travel time formulated and fraction formulated questions are also presented in Sec. IV B. However, a more focused analysis comparing the datasets from the two formulations follows in Sec. IV C 1. In Sec. IV C 2, we investigate potential differences between the two question formulations in more detail. This includes an analysis of the self-assessed certainty scores of the students and the number of guesses made. Moreover, we calculate certain correlations and Cronbach α scores. Correlation in this work is always expressed by means of a Pearson Correlation Coefficient, together with its significance in terms of a p value.

IV. RESULTS

A. Students' estimates

1. Ranking questions

The results of the two ranking questions are summarized in Table I. The ranking of the sizes was considered correct when the Moon was ranked as the smallest and the Earth as the second smallest. The size of the Sun and that of the nearest next star (Proxima Centauri) were considered to be interchangeable, for reasons explained in Sec. III A 1. These ranking questions proved to be more difficult for students than anticipated, as only slightly more than half (53.7%) of the participants succeeded in ranking both the sizes and the distances correctly. Also, ranking the distances was more challenging than ranking the sizes, which resulted in 60.2% of the students ranking the distances correctly and 67.2% ranking the sizes correctly.

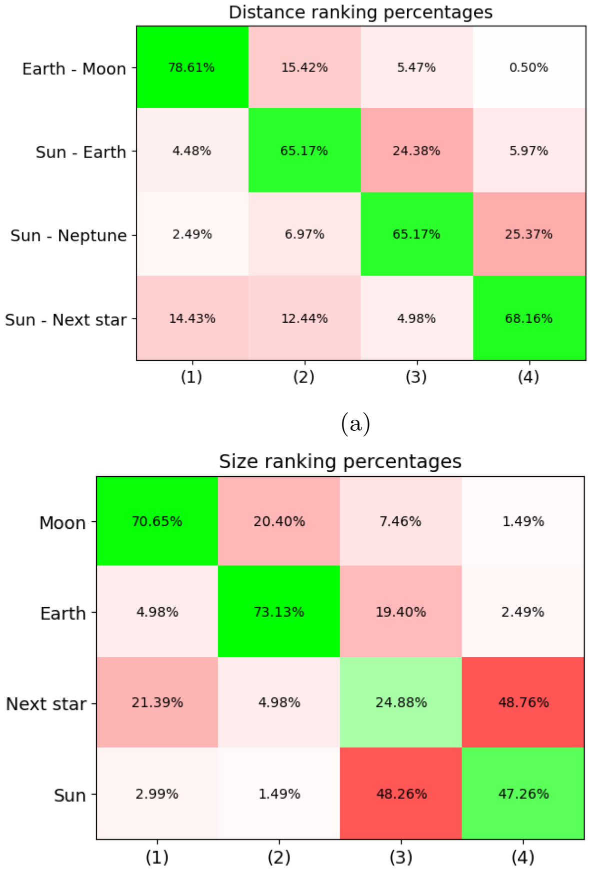
Figure 2 provides a more detailed view of these results. This representation of the results of ranking questions is based on Rajpaul *et al.* [18]. In both panels, the percentage of students ranking a certain distance [in Fig. 2(a)] or size [in Fig. 2(b)] at a certain position in the sequence is displayed for each matrix element. Green and red cells correspond to correct and incorrect answers, respectively, and the degree of saturation matches the magnitude of the percentages. From these figures, it becomes clear that most errors in the ranking questions are related to the nearest next star. On the bottom row of Fig. 2(a), we see that the distance between the Sun and the next star is ranked as either the smallest or second smallest distance in over 25% of the answers. As a consequence, the other three distances are often shifted upward in one place with respect to their correct position, the frequency ranging from over 15% for Earth-Moon to over 25% for Sun-Neptune.

The most apparent feature in the matrix of Fig. 2(b) is the large percentage of students switching the sizes of the Sun and the next star. However, since those two sizes were assumed to be interchangeable, this result should not be seen as disturbing whatsoever. More alarming is the finding that over 20% of students ranked the next star as the smallest of all celestial bodies. Consequently, both the

TABLE I. Main results of the two ranking questions, showing the percentages of students ranking the relevant sizes and distances (in)correctly.

		Sizes		
		Correct (%)	Incorrect (%)	Total (%)
Distances	Correct	53.7	6.5	60.2
	Incorrect	13.4	26.4	39.8
	Total	67.2	32.8	

Moon and the Earth are shifted upward one place in the ranking for very similar percentages of answers. It therefore seems that stars are confusing objects for many students. They imagine stars to be very small objects, of which at least one is situated very close to the Sun. This is in line with the results of Sharp [72], even though his results originate from interviews with much younger children, between 5 and 11 years old.



(a)

(b)

FIG. 2. Results of the ranking question for the relevant distances (a) and sizes (b) in more detail. Each matrix element indicates the percentage of respondents who assigned a specific rank (column number) to a specific item (row). The green cells correspond to correct answers, while the red cells are incorrect rankings. The degree of saturation is added to aid visual clarity, with more saturation corresponding to more prevalent answers.

2. *Quantification questions*

In Fig. 3, the student estimates for the ratio $\frac{\text{Sun-Earth}}{\text{Earth-Moon}}$ are displayed. In this graph, only the students who ranked the Earth-Moon distance as smaller than the Sun-Earth distance —with no other distances in between—were taken into account. This is because, by the way the survey was built, students who failed in that part of the ranking question did not encounter this question further on in the survey. Of course, that does not mean that only the 60.2% of students ranking *all* distances correctly are displayed here; students could e.g., rank the distances of this graph correctly but then rank the Sun-Next star as smaller than both. This particular error was made by 25 students or about 12.5% of the respondents.

In this figure, dots correspond to data from the fraction formulation of this question, and triangles correspond to the travel time formulation. Red data points indicate pure guesses, while estimates are colored according to the assigned certainty. The red horizontal line shows the correct value for this ratio, which is roughly 400. All data shown in this figure were collected after the visualization was shown and the student had agreed with it.

There are some evident observations to be made from this figure. First, a very wide range of magnitudes is estimated by the students, ranging from nearly 1 to 100 000. Also, most students are (very) unsure about their answers, with roughly half of the data (48.7%) corresponding to pure guesses. Third, it can be readily seen that most students underestimate the ratio $\frac{\text{Sun-Earth}}{\text{Earth-Moon}}$.

Similar figures to Fig. 3 can of course be made for all 5 questions of the survey. These plots are shown in Fig. 13, presented in Appendix B. The distributions of all dots in those figures (corresponding to estimates or guesses made for fraction-formulated questions) are represented as boxplots in Fig. 4. The boxplots are placed at the location of their scientific value on the *x* axis. The boxplots to distance-related questions are striped to clearly distinguish them

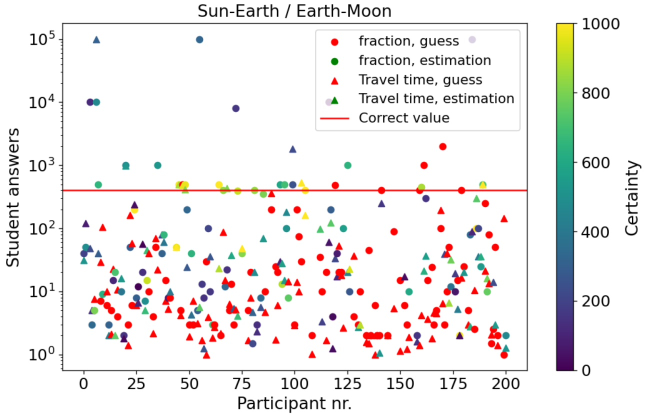


FIG. 3. Student estimates for the ratio $\frac{\text{Sun-Earth}}{\text{Earth-Moon}}$. The *y* axis is plotted on a logarithmical scale, and the certainty of the estimations are visualized through the colormap on the right.

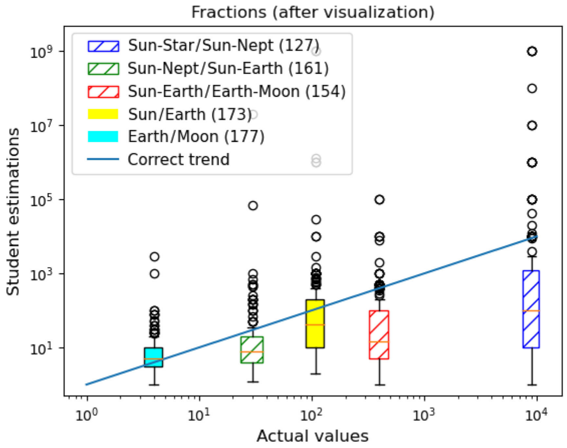


FIG. 4. Distribution of student estimates to all five ratios, originating from the fraction-formulated questions and gathered after approval of the visualization.

from the fully colored boxplots for the size-related questions. In the legend, the number of students that answered the question of that ratio is mentioned. That number varied for every question because it depends on the results of the ranking questions, as explained earlier.

From this figure, it is readily observed how all relative distances in this survey were strongly underestimated by most students. The same does not hold for the two questions on relative sizes, where we see more comparable amounts of both over- and underestimation. Also, the wide spread of the estimates appears to be a recurring characteristic.

B. Effect of visualizations

Figure 5 represents students' answers before the visualizations were shown. There is clearly a lot of similarity between Figs. 4 and 5. The distance ratios are still underestimated, and there is still a wide range of estimates for all questions. However, there are also some differences to be recognized. For example, both boxplots related to size questions seem to shrink toward the lower end after visualization. This means that after seeing the visualization with the two disks, several participants judged the contrast in their sizes too large and reduced the initial answer.

The limited visible difference between Figs. 4 and 5 does not indicate that no adjustments were made after the visualizations. Table II shows the number of students who adjusted their answers to each question. To display these data for all questions individually, the results for fraction and travel time questions must inevitably be presented separately. However, the analysis comparing in detail the results between the two question formulations only follows in Sec. IV C.

Overall, just under one-third of the answers were changed, both in the fraction formulation and in the travel time formulation. Furthermore, the percentages of adjusted answers are very similar between the two question formulations on the level of individual questions as well.

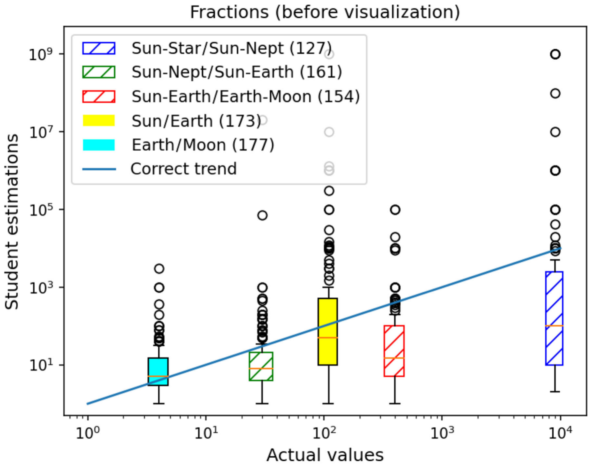


FIG. 5. Same distribution of student estimates as Fig. 4, but for data collected before any visualization had been shown.

However, there is some difference in the percentage of adjustments between distance-related and size-related questions. For the three distance-related questions combined, roughly 25% of the answers were adjusted, while for the two size-related questions this was 40% (for both question formulations). Specifically for the question on the ratio $\frac{\text{Sun}}{\text{Earth}}$, several students initially gave a relatively high answer, but lowered their answer after seeing the two disks with very contrasting sizes. This effect is more clearly illustrated in Fig. 6. Note that in this figure, only the data from the 78 students who altered their answers to this question in the fraction formulation are shown.

As mentioned in Sec. III, half of the participants answered the fraction formulated questions first and the travel time questions second, while the other half worked in the opposite order. The numbers of adjustments were compared between the participants of both groups, for each question individually. This was done using a

TABLE II. Numbers and percentages of students adjusting their answer after seeing a visualization, for all questions and both question formulations.

	Fractions	Travel times
Sun-Star/Sun-Neptune	39/127 30.7%	34/127 26.8%
Sun-Neptune/Sun-Earth	40/161 24.8%	42/161 26.1%
Sun-Earth/Earth-Moon	31/154 20.1%	34/154 22.1%
Sun/Earth	78/173 45.1%	73/173 42.2%
Earth/Moon	63/177 35.6%	67/177 37.9%
Total	251/792 31.7%	250/792 31.6%

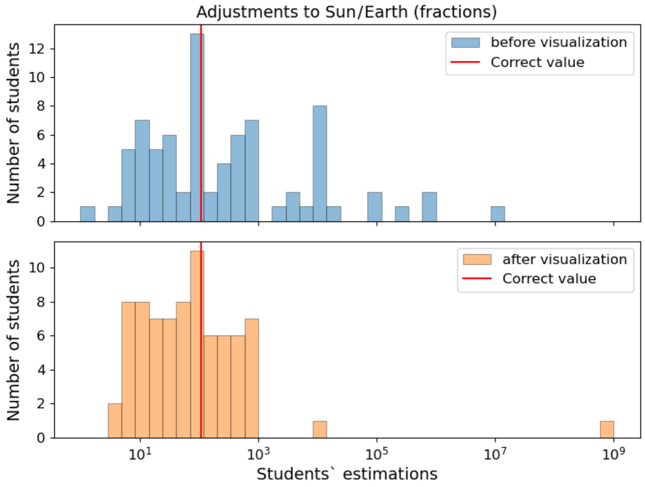


FIG. 6. Initial and final estimates for the ratio $\frac{\text{Sun}}{\text{Earth}}$, given by the 78 students who altered their initial answers to this question in the fraction formulation after seeing the visualization.

two-proportion z test, where adjustments were considered as “successes” and the participants in each group as “tries.” For all but two questions, the difference in the number of adjustments was not significant, with $p > 0.05$. Only for the questions probing $\frac{\text{Sun-Star}}{\text{Sun-Neptune}}$ and $\frac{\text{Earth}}{\text{Moon}}$, both in the fraction formulation, we found weakly significant differences with $p \approx 0.044$ and $p \approx 0.040$, respectively. In both cases, the proportion of students making adjustments was larger in the group that treated the fraction questions first. We therefore interpret that there was some effect of fatigue on the number of adjustments, but it was very limited overall.

A logical question to ask now is whether these adjustments after visualization bring students closer to the correct answers. The results answering this question are written in Table III. It is important to note that in the analysis leading to these results, correctness was evaluated as the logarithmic offset of the student’s answer to the correct value. For example, for the ratio $\frac{\text{Sun-Earth}}{\text{Earth-Moon}}$, a factor of 2000 is considered more correct than a factor of 40, since $\log(\frac{2000}{400}) < \log(\frac{400}{40})$. For completeness, we should add that the use of “log” throughout this paper always refers to a logarithm of base 10. Also indicated in Table III are the levels of significance. These levels are calculated using a two-sided binomial test, assuming the probability of making an adjustment that improves the correctness of the answer to be 50% under the null hypothesis. This is not a trivial assumption, since the probability of improving the initial answer depends on how close that answer is to the correct value. However, we argue that this 50% can be taken as a high-end limit, resulting in corresponding high-end limits for the p values. The actual probability of improving an initial answer is lower than 50%, since the students first had to adjust in the correct direction (increasing or decreasing their answer), and subsequently they had to prevent overshooting the correct value by such an amount that the answer would become more incorrect. Choosing the right direction has a 50-50 chance, and the prevention of overshooting further decreases the chance of improving an initial answer. The binomial test was chosen because it provides a simple yet robust way to evaluate the effect of the adjustments and to decide on the statistical significance of those effects.

Showing a visualization did not lead to a significantly worse answer for the group of students who made adjustments to any of the questions. This can be deduced from the fact that all statistically significant results in Table III correspond to percentages above 50%, meaning that the majority of students have improved their initial answers. However, of course, for all questions, there were individual students who worsened their initial answers by making the adjustment (for the Sun-Star/Sun-Neptune question in the fraction formulation, this was even the case for 25 out of 39 students, which has $p \approx 0.1$). Collectively speaking, adjusting their answers therefore either had a zero net effect (meaning there was no statistical significance) or improved the correctness of the answers. For distance-related questions, visualizations appeared to be more effective in improving initial answers within the travel time formulation. For size-related questions, the visualizations had a greater effect on the fraction formulation. Also note that for all questions combined, the beneficial effect of the visualizations is moderately or even very significant for fraction and travel time questions, respectively.

A possible remark to these results is that by using binomial testing, only the *number* of improved answers is taken into account, and not the *amount* by which an answer is improved or worsened. To counter this caveat, a parallel analysis was performed while using Wilcoxon tests. This test was preferred because the data before and after visualization can be considered paired, and the nonparametric nature of

TABLE III. Numbers and percentages of adjusted answers that improve the correctness, for all questions. Significance levels: $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Fractions	Travel times
Sun-Star/Sun-Neptune	14/39 35.9%	28/34*** 82.4%
Sun-Neptune/Sun-Earth	18/40 45.0%	36/42*** 85.7%
Sun-Earth/Earth-Moon	21/31 67.7%	23/34 67.6%
Sun/Earth	50/78* 64.1%	46/73* 63.0%
Earth/Moon	46/63*** 73.0%	39/67 58.2%
Total	149/251** 59.6%	172/250*** 68.8%

the test rules out any dominance of extreme outliers. The resulting p values from the Wilcoxon tests were at least as significant as the ones in Table III for every question, and occasionally even more significant.

C. Effect of question formulation

1. (In)consistency

While some students were very consistent in answering the same question over two formulations, others gave very different estimates. Moreover, they often assigned similar self-assessments of certainty to both answers. This could already be noticed by looking more closely at Fig. 3, where the dot and triangle for a certain student were in some cases separated by multiple orders of magnitude. The finding that these individual inconsistencies are no exception is even more clearly shown in Fig. 7. This figure shows the correlation between the correctness of a certain answer to a question in the fraction formulation and the correctness of the answer to the same question in the travel time formulation. As before, the correctness of an answer is evaluated in a logarithmic manner, calculated as $\log(\frac{\text{Student answer}}{\text{Correct value}})$. The data in this figure show the answers of all students, for all questions combined.

Data points on the red line are very consistent answers, in the sense that they over- or underestimate a certain ratio by the same amount in the two question formulations. Despite multiple data points lying very far from this red diagonal—and thus representing very inconsistent answers—the correlation of this data is significant ($p < 0.001$) and large ($R \approx 0.64$). Therefore, students who tend to overestimate a fraction-formulated question will generally also overestimate the corresponding travel time question, and analogously for students underestimating or giving accurate answers. However, the figure also shows that there are many more data points situated under the red diagonal than above, more specifically this is the case for 66.9% of the data. For all those data, the considered ratio was estimated to be lower in the travel time formulation than in the fraction formulation. This proved to be a general trend, present for all question pairs. The strength of this trend for every question individually is displayed in Table IV. The significance levels are calculated by means of a binomial test, assuming that the probability of a fraction answer being larger than a travel time answer is 50% under the null hypothesis. It is clear from Table IV that questions in a fraction formulation lead to higher estimates than questions in a travel time formulation. Given that the medians of the answer sets to all ratios except Earth/Moon were clearly below the correct value (see Figs. 4 and 5), this implies that fraction formulations may also lead to more correct answers. To verify this, we calculated the number of fraction—travel time pairs in which the fraction answer was more accurate (again, in a logarithmic sense). The fraction answers indeed were found to be more correct than the travel time answers, with significance levels of at least $p < 0.05$ for all ratios except the Sun-Earth/Earth-Moon. The significance levels were stronger for the remaining two distance-related questions than for the two size-related questions, because for the distance-related questions, there was much more underestimation.

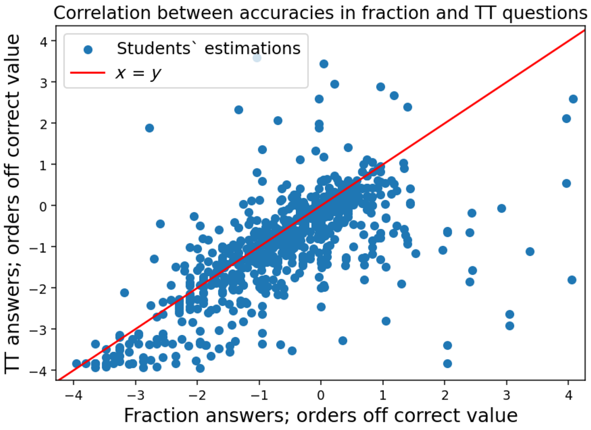


FIG. 7. Correlation between the correctness of an answer in the fraction formulation and the correctness of the answer to the corresponding question in the travel time formulation, both evaluated after visualization.

TABLE IV. Number of answer pairs where the answer in the fraction formulation was larger than the corresponding answer in the travel time formulation. Significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

	Before visualization	After visualization
Sun-Star/Sun-Neptune	101/127*** 79.5%	98/127*** 77.2%
Sun-Neptune/Sun-Earth	117/161*** 72.7%	109/161*** 69.6%
Sun-Earth/Earth-Moon	96/154** 62.3%	95/154* 60.4%
Sun/Earth	108/173** 62.4%	116/173*** 67.1%
Earth/Moon	119/177*** 67.2%	112/177*** 63.3%
Total	541/792*** 68.3%	530/792*** 66.9%

Similar to the results in Tables III and IV only shows that a significant *number* of answers was higher in the travel time formulation, but not necessarily that the two datasets differed by a significant *amount*. Since the answers of the same student are compared between the two formulations, and since there are definitely extreme outliers in the data, the Wilcoxon test was once again used to complement the previous results. The resulting significance levels were very similar to the ones found with binomial testing, the only difference being an occasional change from $p < 0.05$ to $p < 0.01$ or vice versa.

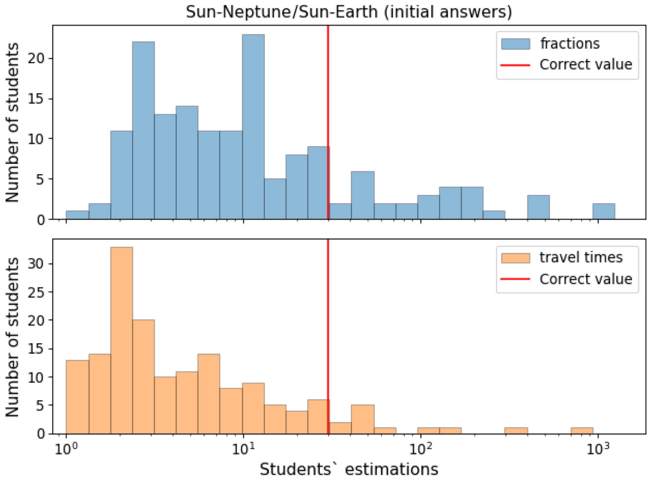


FIG. 8. Example of a question where a fraction formulation clearly leads to higher estimates than a travel time formulation.

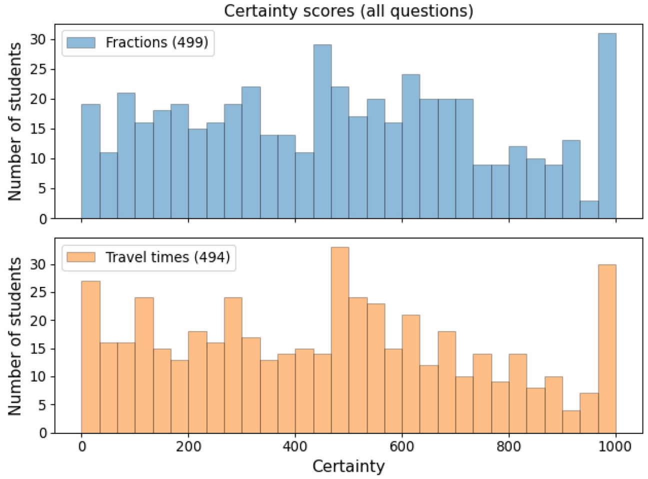


FIG. 9. Distribution of the self-assessed certainty of the students, for both question formulations and all ratios combined. The legend mentions the amount of answers with an assigned certainty score, that is, the amount of nonguesses.

Once again, it should be checked whether the order of questions had any influence on the results of this section. This was evaluated using a Mann-Whitney U test, comparing the answers to all ten questions separately (five in each formulation) between the participants answering fractions first and those answering travel times first. The Mann-Whitney U test was chosen because it is a nonparametric test (so that the effects of extreme outliers are neutralized) and because the compared sets of data were not paired. Regarding the answers before visualization, there were three questions in which the datasets differed significantly, with $p < 0.05$. After visualization, this decreased to only one question (on the ratio $\frac{\text{Earth}}{\text{Moon}}$ in the travel time formulation). Overall, the magnitudes of students' answers did therefore not nearly depend as much on question order as they did on question formulation.

In Fig. 8, a clear example of this trend is shown. In the discussion, we will elaborate on a possible reason for this observation. More important now are its implications; if one question formulation leads to systematically different estimates than another, which set of answers should we then take to represent best the students' mental models? In the next section, we seek indications in the data that one formulation should be preferred over the other.

2. *Search for a question formulation preference*

As a first note to this section, we highlight that both question formulations were already demonstrated to be mutually compatible, to a satisfactory degree. To substantiate this claim, we refer to Fig. 7, where an overestimation (underestimation) to one question formulation was shown to clearly correlate with a similar overestimation (underestimation) to the other question formulation. If this were not the case, then the two question formulations would be probing completely different things, and our research objective to assess student mental models of astronomical scales would essentially be jeopardized. However, despite the correlation between answers to both question formulations, there could still be indications in the data that one formulation would be potentially preferable over the other.

As a first indication, the self-assessed certainty scores were analyzed. Should students evince greater certainty in their responses to one question formulation than to the other, we expect that their answers to the former are more reliable. Similarly, a smaller number of "pure guess" answers for one question formulation would also increase its reliability. However, no such differences were found between the two question formulations, as illustrated by Fig. 9. The answers to fraction and travel time formulated questions were assigned very similar certainty scores overall, as shown using both a t test and a Mann-Whitney U test. These tests were preferred over the previously mentioned binomial and Wilcoxon tests, as now the two datasets could no longer be considered paired (because the students making certainty assessments for a question in one formulation could make a pure guess for the same question in the other formulation). Both the t test and the Mann-Whitney U test resulted in a nonsignificant difference between the certainty scores in the two formulations, with $p > 0.05$. Moreover, as can be inferred from the legend in Fig. 9, an almost equal number of pure guesses was made in the two formulations.

The certainty scores were also compared to the correctness of the corresponding answers, for both question formulations. If one formulation would result in a much higher correlation between the certainty and the correctness of the answers than the other formulation, that would indicate a higher accuracy in the self-assessment of one's knowledge in the former formulation. However, as shown in Fig. 10, only very small correlations between these

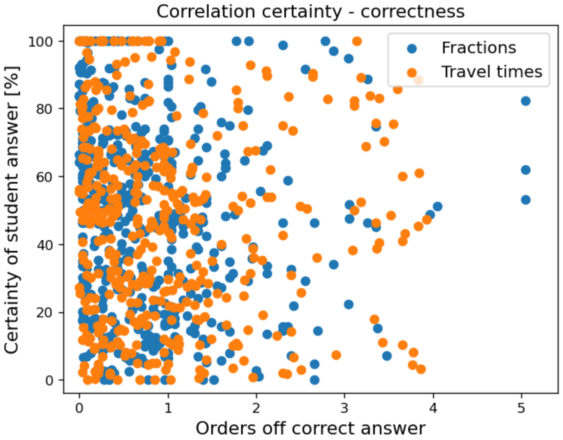


FIG. 10. Correlation plot between the self-assessed certainty of the student answers and the corresponding correctness, which is expressed in terms of the orders of magnitude separating the answer from the correct value. The data are shown for both question formulations and for all ratios combined.

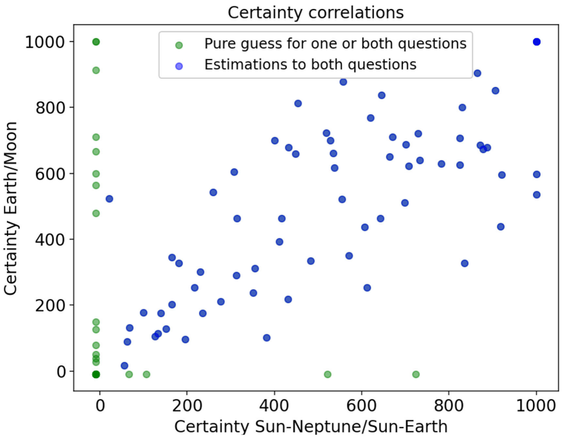


FIG. 11. Correlation plot between the self-assessed certainties to the questions on the ratios $\frac{\text{Earth}}{\text{Moon}}$ and $\frac{\text{Sun-Neptune}}{\text{Sun-Earth}}$, both in the fraction formulation. Pure guesses to one or both questions were assigned a certainty of -10.

variables were found ($R \approx -0.19$ for fractions and $R \approx -0.07$ for travel times), with weak or no significance ($p \approx 0.021$ for fractions and $p \approx 0.14$ for travel times). This was much to our surprise, as it was anticipated that students would generally be more certain about more correct answers.

While certainty scores were found not to correlate with the accuracy of the answers, they did correlate over several questions. That is, the certainty scores students gave to two questions were found to be correlated, to varying degrees. An example of this result is shown in Fig. 11, comparing the certainties to the fraction-formulated questions on the ratios $\frac{\text{Earth}}{\text{Moon}}$ and $\frac{\text{Sun-Neptune}}{\text{Sun-Earth}}$. Students who indicated that at least one of their two answers was a pure guess are indicated by green data points, while students who did not guess either answer are shown in blue. Pure guesses are shown at a certainty value of -10. The correlation in this example is significant with $p < 0.001$ both with and without the green data points taken along in the calculation. We found $R \approx 0.72$ when only considering the blue data points, and $R \approx 0.68$ while also considering the green data points. These values only slightly changed when altering the certainty value of a pure guess, for example from -10 to -1000. We conclude from the results of Fig. 11 that students answering with great certainty to one question are likely to also be certain about other answers, while Fig. 10 illustrates that such high certainties are not at all indicative of more accurate answers.

A second possible criterion for a preference in question formulation could be the number of adjustments made after seeing a visualization. As explained in Sec. IV B, more adjustments could point out that students find it harder to align their answers with their mental model, therefore making that formulation less suitable for administration. However, this hypothesis was already debunked by the results of Table II, where the numbers of adjustments were shown to be very similar between the two formulations.

What Table II does not show is the extent to which the students adjusted their answers. If the adjustments made for, e.g., fraction questions were significantly larger than those made for travel time questions, this would imply that students have more difficulty aligning their (initial) answers with their mental models when answering fraction questions. In Fig. 12, the magnitudes of the performed adjustments for both question formulations are shown. The magnitudes are expressed as $\frac{\text{Adjusted answer}-\text{Initial answer}}{\min(\text{Adjusted answer,Initial answer})}$. The denominator of this formula represents the smallest of the two given answers. This denominator was chosen to obtain a relative measure of the correction magnitudes, while maintaining an equal weight for increases and reductions of initial answers. Note that with this formula, a correction magnitude of 1 (-1) corresponds to a doubling (halving) of the initial answer. Analogously, a magnitude of 10 (-10) indicates an increase (reduction) by a factor 11, ±100 by a factor 101, and so on.

We observed no significant difference in adjustment magnitudes between the two panels of Fig. 12. As can be seen from the two legends, there were slightly more reductions in initial answers to fraction questions and more increases in answers to travel time questions, but the two medians lie comfortably in each others' interquartile ranges.

Finally, we explored the internal consistency of the answers students gave across the test to both question formulations. The objective was to investigate whether students answering (relatively) correctly on one question would also be likely to answer well on other questions of the same formulation. This is in some way analogous to what is shown in Fig. 7, where we investigated a correlation between the correctness of two identical questions in different formulations. Now, however, we take into account

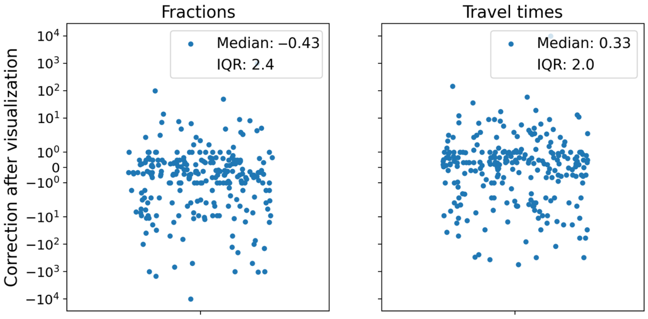


FIG. 12. Magnitudes of the answer adjustments, for all questions combined. This magnitude is expressed as $\frac{\text{Adjusted answer}-\text{Initial answer}}{\min(\text{Adjusted answer, Initial answer})}$.

TABLE V. Cronbach α values and EV scores of the first component of PCA for both question formulations. Evaluations have been done before and after visualizations, and with either absolute values or signed values of the errors.

		Fractions	Travel times
Before visualization	|Error|	$\alpha = 0.27$	$\alpha = 0.42$
		EV = 0.34	EV = 0.35
	±Error	$\alpha = 0.58$	$\alpha = 0.57$
		EV = 0.41	EV = 0.39
After visualization	|Error|	$\alpha = 0.36$	$\alpha = 0.51$
		EV = 0.34	EV = 0.38
	±Error	$\alpha = 0.58$	$\alpha = 0.64$
		EV = 0.40	EV = 0.44

five different questions of the same formulation and compare the results between the two formulations. As before, we evaluated the correctness of every answer in a logarithmic way to limit the effect of extreme outliers. Subsequently, we computed the Cronbach α score as a measure for internal consistency between the correctness of all answers of a particular student. Additionally, we performed a principal component analysis (PCA) to obtain the explained variance (EV) of the first principal component. Those two variables provide related, though slightly different information. Cronbach α represents the true internal consistency, extracted from the correlations between all sets of student answers. A higher Cronbach α score would indicate a better reliability of the whole set of questions. The EV of the first principal component, on the other hand, expresses how "aligned" the student answers are in 5D space, thereby providing more insight into the structure of the data. By calculating both variables, we aim to gain a more complete view of the coherence level of the datasets. While this part of the analysis may be considered rather far-reaching, it is motivated mainly by the previous nonsignificant results. The calculation of these variables was deemed worthwhile to pursue because, as no formulation preference was uncovered in terms of previously considered data clues, a discrimination criterion was still to be found.

The results of this analysis are shown in Table V. Results are displayed for the answers submitted both before and after visualization. Moreover, we also evaluate these variables while using both the absolute values of the errors (that is, $|\log(\frac{\text{Student answer}}{\text{Correct value}})|$) and the signed values.

Although there are some differences in these results between the two question formulations, those are insufficient to cause an unambiguous preference for one question formulation. The conclusion of this section is therefore that, despite the fact that the two question formulations result in measurably different estimates, we have found no indication that one formulation is more closely probing students' mental models than the other.

V. DISCUSSION AND FUTURE RESEARCH

In this work, we constructed and administered an interactive, online survey to 201 students in their last year of high school. Their estimates of three relative distances and two relative sizes of astronomical objects in the Solar neighborhood were assessed. Two question formulations were used: one numerical comparison which we named the "fraction" formulation, and one formulation in terms of an imaginary spacecraft which was named the "travel time" formulation. The differences between the results of those two formulations formed one of the main interests of this study. Customized visualizations were shown after every answer, with which students could either agree and move on to the next question, or disagree and reconsider their previous answer until they agreed with the visualization. Additionally, self-assessments of certainty were collected for every answer.

The first clear result of this study was that students in general drastically underestimated all assessed distance ratios. As we have only questioned these distances in a relative sense, it would not be very accurate to state that students underestimate these distances per se, but rather that they underestimate the leaps in distance between subsequent celestial bodies. For example, we have not actually found that students underestimate the distance from Earth to the Sun, but rather that they underestimate how much farther the Sun is from Earth compared to the Moon. More specifically, from Figs. 4 and 5, it can be deduced that these underestimations become more severe as the relative distance to be estimated becomes larger. Additionally, the spread in answers also becomes wider with increasing distance, most clearly seen by comparing the interquartile ranges of the (striped) boxplots. These results are in strong agreement with those of, e.g., Miller and Brewer [17]. In future work, it would be interesting to investigate whether this trend continues for even larger distances, for example those between galaxies.

For the relative size questions, the observed trend is less pronounced. The Earth/Moon ratio is estimated quite well by most participants, and the range of magnitudes of the answers is relatively small. TheSun/Earth ratio has a larger range of answers, and the number of students underestimating this ratio starts to dominate over those overestimating. However, this dominance is not strong enough to conclude that this ratio is underestimated as much as the ones related to distances. We suspect that the degree of underestimation would continue to grow when larger ratios are assessed. After all, the range of magnitudes for the probed relative sizes (up to roughly 100) was a lot smaller than that for the relative distances (up to 9000). The more moderate results for relative sizes may be a consequence of this smaller magnitude span. Similar to distances, it would be interesting to investigate in future research how this trend evolves for larger sizes, for example, that of the Milky Way. Of course, such questions could only be asked to participants who are familiar with these structures, which presumably does not hold for all secondary school students.

It is doubtful whether a canonical size as described by Konkle and Olivia [62] (that is, a preferred size at which an object is drawn or imagined, see Sec. II B 2) also manifests itself for astronomical objects. Since their study concerned objects that "can be viewed at a range of distances and thus can be experienced at a range of visual angles within the visual field," celestial bodies are plausibly too large to fit into this framework. The great range of size estimates made by the students, and therefore also the range of sizes seen in the visualizations with the two red balls, seem to confirm this point.

Of all answers, roughly half were indicated to be pure guesses. The assigned certainties of the nonguesses were very diverse, with an approximately uniform distribution for both fractions and travel times (see Fig. 9). These findings, together with the broad range of student estimates, suggest that many participants really struggled to answer the questions in our survey. Clearly, many students simply did not know the sizes and distances involved in our Solar System, in which cases major inconsistencies between fraction and travel time answers are hardly surprising. This lack of knowledge was already clear in the results of the ranking questions, with only 54% of the students ranking both distances and sizes correctly. However, students are not necessarily aware of their astronomical ignorance, as shown by the absence of a strong correlation between their self-assessed certainty and the correctness of their answers. We therefore fully agree with Lelliott and Rollnick's [8] previously mentioned statement that the topic of distances and sizes in astronomy is undertaught. However, we should add here that a substantial amount of recent research is being devoted to teaching methods that improve students' understanding of this topic. Efforts made include the use of virtual simulations [1,41], zooming videos [74], playing cards ranked by size [9], and planetarium visualizations [75], with varying but overall promising results.

We found that the exact question formulation has a measurable impact on student estimates of astronomical scales. Answers originating from the fraction formulation were higher than those from the travel time formulation, for a significant percentage of the students and for all assessed ratios. Moreover, many students were inconsistent between their answers to similar questions in the two formulations, with occasionally several orders of magnitude separating the two estimates of the same ratio. Future researchers probing estimates on astronomical scales should be aware of the (subtle) impact that question formulations may have on the study outcomes. Similar research probing estimates on microscopic, atomic, or subatomic scales should consider similar effects to potentially arise in these fields as well. Coordinating assessment methods over different studies could possibly avoid or reduce disagreements between the results and conclusions of those studies.

At the time of writing, we have no substantiated explanation for the higher estimates in the fraction formulation than in the travel time counterparts. It appears likely that the discrepancy results from different mental operations made by the students while answering both question types. Based on observations during previously performed pilot interviews, we hypothesize that students are more reluctant to make large "jumps" in travel times than they are to state large fractions. Inversely, students find it easier to answer with a vast number for a ratio they know to be large than to provide a similarly large ratio in terms of a travel time. For example, one participant of the online survey stated that the distance Sun-Neptune is 100 000 times smaller than the distance Sun-Next star, probably because he knows the ratio to be large and 100 000 is definitely a large number. However, when the latter distance is said to take 4 years of travel time, this student estimated the former distance to take five days, probably because he also rates that as a large contrast in magnitudes. However, 5 days is only (roughly) 300 times smaller than 4 years. This student indicated both answers to be nonguesses, with a certainty of 46% and 12% for the fraction and travel time questions, respectively.

Despite the measurable difference in estimates between the two question formulations, we have found no indication that the results of one formulation better represent the student mental models than the results from the other formulation. A comparable number of guesses was made for both formulations, the results of the self-assessments of certainty were similar, and students behaved no differently between the two formulations when interacting with the visualizations. No differences were found in the internal consistency of both sets of answers, and none of the formulations resulted in a meaningful correlation between the certainty of the answers and their correctness. As such, we can give little suggestions on how questions on this topic should or should not be formulated in future research, except for the remarks made in Sec. II A 3. Rather, we must conclude that the expressed models of astronomical scales in the Solar neighborhood are found to be subject to the external factor of question formulation, but that both formulations lead to equally valid answers. In future research, the inclusion of a third question formulation might shed new light on this topic.

None of the above findings can be attributed to the order in which the questions were asked. For most questions, there was no statistically significant impact of question order on either the magnitudes of the estimates or on the number of adjustments after visualization. For the questions where a difference was found, the p value was generally only slightly below the significance threshold ($p < 0.05$). We therefore neglect any potential effect of fatigue in this discussion.

The visualizations caused roughly one-third of all answers to be adjusted. This number was measurably higher for the size-related questions than for the distance-related questions. This could indicate that size-related questions were perceived as more difficult so that students often needed more than one attempt to align their answers with their mental model. Alternatively, it could also mean that the visualizations for the size questions (with the two-dimensional disks) were more immersive and therefore more likely to trigger second thoughts than the visualizations for distance questions (which were essentially one-dimensional lines). Additionally, it should be pointed out that areas are perceived differently than lengths [76], and viewers can generally state the ratio of two lengths more precisely than that of two areas. Although student answers to size-related questions are encoded as one-dimensional radii, the ratio of the visualization is perceived in terms of the areas of the discs. Even when the linear ratio is explicitly mentioned, this could lead to a "deception of message exaggeration/understatement" [77]. The discrepancy between the linear answering format and the squared perception of the visualizations may be another cause for the more numerous adjustments in size-related questions.

For some but not all questions, a significant percentage of the adjusted answers were improvements (see Table III). This effect seemed to be more strongly present in the travel time formulation for distance-related questions and in the fraction formulation for size-related questions. For those questions where the improvement effect was significant, one could argue that providing students with the opportunity to interact with a visualization has a positive effect on the results. Moreover, not including any visualizations in those questions entails the risk of drawing too pessimistic conclusions about the student's knowledge. However, also in those cases where the adjustments of answers were not leading to significantly better estimates, there are reasons to believe that the adjusted answers more closely represent the students' mental models (as elaborated on in Sec. II B 2). Therefore, we believe these customized representations to have an added value to the survey, regardless of whether the adjustments improved student answers and regardless of the question formulation.

With two-thirds of all answers not being adjusted after seeing the visualizations, we should not be overly enthusiastic about its impact. For all those unaltered answers, it is open to discussion whether the students immediately agreed with the visualizations because there was an instant match between what they were seeing and their mental model. Alternatively, they could have agreed out of a lack of interest in the survey, or because they simply did not know the answer. From the data obtained through our study, it is difficult to discriminate between these motives. Most probably they all contributed to the results, but their relative importance remains unknown. What further hindered the gain of insight into this matter was the absence of any correlation between the correctness of students' answers and the assigned certainties, as shown in Fig. 10. Highly accurate but uncertain answers could either originate from a lack of confidence or from lucky guesses (although these answers were not indicated to be pure guesses). Similarly, answers with low accuracy but high certainty could represent real misconceptions or display a general overconfidence among some students. Indeed, varying degrees of correlation were found between the self-assessed certainty of students in two different questions, an example of which is displayed in Fig. 11. This contributes to the interpretation that these self-assessments of certainty tell us more about the students themselves and their overall self-confidence than about the correctness of their answers.

As argued by diSessa [78], p prims are fragmented in the sense that they give rise to different responses to questions that are fundamentally similar but differ in superficial characteristics, e.g., they are phrased in different ways. At first thought, one could thus argue that the different answers resulting from different question formulations indicate that internal knowledge structures are organized more like an assembly of fragmented units of knowledge than as a fully coherent framework theory. However, supporting or falsifying any of the theories on knowledge structure was not the aim of this work, and we will restrain from taking any strong position in the debate. Nonetheless, there are some important remarks to make on this account. First, as Fig. 7 illustrates, most students were only mildly inconsistent in their two answers to the same question. The alignment of the data around the red diagonal line in the figure, together with the large R and low p values, illustrate how the majority of the answer pairs differ by less than one order of magnitude. The question could be asked, then, by how much do two answers need to differ in order to be considered inconsistent?

A second remark to highlight is that our results merely show how different question formulations trigger different *expressed models* of the participating students. To extrapolate this difference to conclusions about the underlying *mental models* of the students is nontrivial. It is possible that the mental models of the students are in fact stable, but their expression is subject to the interrogation method. While our survey was thought to be well constructed and efficient, it generates purely quantitative data, which probes the student models rather superficially. Further exploring this matter and digging deeper into the nature and (in) stability of these models requires more qualitative data in the form of e.g., verbal interviews.

VI. CONCLUSION

Assessing estimates on astronomical scales remains a nontrivial research objective. Not only do students (but presumably not only students) struggle with making such estimates, but the precise methodology also affects their answers. In this study, we have shown this effect by demonstrating how different question formulations lead to measurably different answers. The magnitude span of the estimates covered many orders of magnitudes for all questions in both formulations, which complicated the statistical analysis. However, regardless of which formulation was used, we found that students systematically underestimate the vastness of space, which is in agreement with many previous studies (see Sec. II for references). As these students' estimates were found to be (to varying degrees) inconsistent over the course of our survey, future research should dedicate time and effort to exploring how we can further develop such inherently unstable expressed models. The inclusion of customized visualizations is considered a valuable step in the right direction but did not really lift the discrepancy in answers between the two question formulations. In order to gain insight into this topic, it is essential to uncover the reasons why students give certain answers. For instance, does a student agree with a visualization because he has limited enthusiasm for the test or because the image truly matches his mental model? Qualitative research is essential to better grasp students' reasoning while making these judgments. Individual interviews are necessary to extract an in-depth view of the students' mental models on the scale of the Solar System.

ACKNOWLEDGMENTS

The authors would like to thank Timon Vanslambrouck from the IT user workplace support team of KU Leuven for his help in constructing the JavaScript code behind the interactive survey. This work was financially supported by KU Leuven (STG/21/039).

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because of legal restrictions preventing unrestricted public distribution. The data are available from the authors upon reasonable request.

APPENDIX A: PILOT STUDIES

Prior to the administration of the online survey presented in this paper, two rounds of pilot interviews were performed. In the first and second rounds, there were 17 and 14 participating students, respectively. These students were all in their last year of high school at the time of the interview and had the same (limited) astronomy instruction at school as the participants in the online survey. In both rounds of interviews, the addressed relative sizes and distances were the same as the ones in the main study, as listed in Sec. III A. Also in both rounds, participants were asked to rank these sizes and distances in increasing order at the start of the interview.

In the first round of pilot studies, the distance-related questions were asked in the same travel time formulation as discussed in Sec. III A 1, while the size-related questions were asked by referring to a small-scale model of the Moon, Earth, and Sun. An example of such a size-related question was: "If we had a small-scale model of the Solar System, and on this scale the Sun would have a diameter of one meter, then what would be the diameter of the Earth?" All these questions would be asked twice to each student. In the first round of questions, we started out with the smallest size (distance) and systematically worked toward larger sizes (distances), while in the second round, we worked in the opposite direction. After each question, students were presented with a visualization of their previous answer. These visualizations were very similar to what was shown in the online survey (as illustrated in Fig. 1) but now drawn manually on paper by the researcher.

In addition to these questions on astronomical scales, the students in the first piloting round were also asked to complete 23 bounded NLE tasks. The main goal of including these questions was to investigate whether the participants could correctly position the relevant ratios on a line. Therefore, the correct values of all relevant ratios (being approximately 4 and 100 for the size questions, and 30, 400, and 9000 for the distance questions) were all included multiple times in the NLE tasks. The choice for bounded over unbounded NLE tasks was a rather pragmatic one since unbounded NLE tasks were deemed unfeasible with such large values to be assessed.

In the second round of pilot interviews, both the distance-related and size-related questions were first asked in a fraction formulation and subsequently in a travel time formulation. This is identical to the questioning formulations applied in the online survey later on, as addressed in Sec. III A 1. The change in questioning format with respect to the first piloting round was made to create a closer parallel between the distance-related and size-related questions. All questions were now asked by starting from the largest size/distance and working toward the smallest one, a choice motivated by the rather accurate performance of the students in the NLE tasks during the first piloting round. No more NLE tasks were asked in the second round of interviews. Instead, a self-assessment of certainty was filled in by the students after every question in the form of a 1–4 Likert scale, including a “pure guess” option. By differentiating between random guesses and (un)certain estimations, we could obtain a more detailed idea of the student mental models, which was more difficult with the data obtained through the first piloting round. After this second round of interviews, we felt confident enough to enter the main study phase by questioning a larger number of students. In order to do so in a time-efficient way, we decided to move from individual interviews to an online survey.

APPENDIX B: ADDITIONAL FIGURES

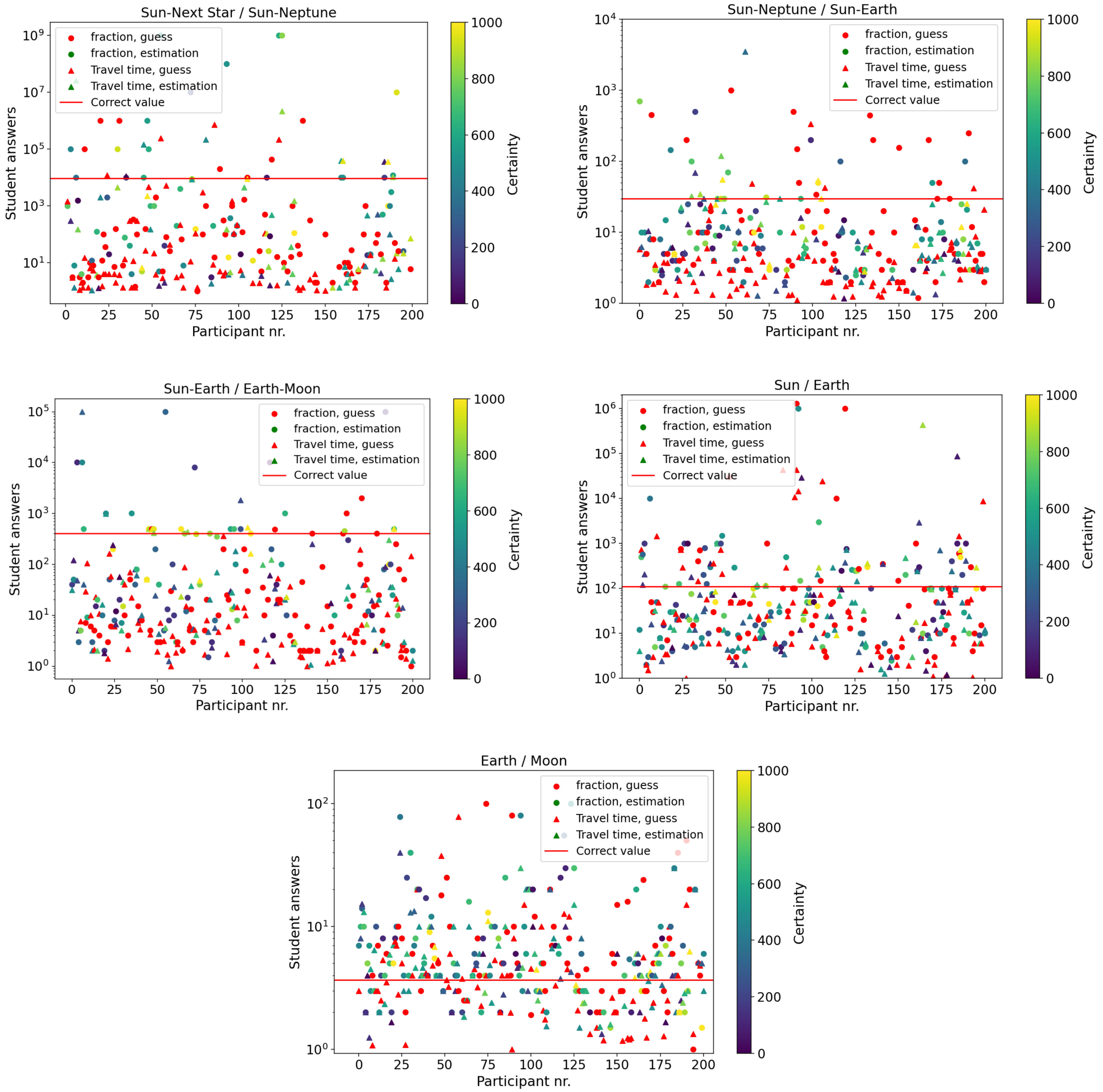


FIG. 13. Results of all quantification questions.

[1] M. Schneps, J. Ruel, G. Sonnert, M. Dussault, M. Griffin, and P. Sadler, Conceptualizing astronomical scale: Virtual simulations on handheld tablet computers reverse misconceptions, Comput. Educ. **70**, 269 (2014).

[2] D. Landy, N. Silbert, and A. Goldin, Getting off at the end of the line: The estimation of large numbers, in *Proceedings of the Annual Meeting of the Cognitive Science Society* (Cognitive Science Society, Austin, TX, 2012), Vol. 34.

[3] D. Landy, A. Charlesworth, and E. Ottmar, Categories of large numbers in line estimation, Cogn. Sci. **41**, 326 (2017).

[4] A. Brass and S. Harkness, Pre-service teachers' conceptions of the magnitude of large numbers, Invest. Math. Learn. **9**, 53 (2017).

[5] S. Kastberg and V. Walker, Insights into our understandings of large numbers, Teach. Child. Math. **14**, 530 (2008).

[6] J. M. Bailey, K. Coble, G. Cochran, D. Larrieu, R. Sanchez, and L. R. Cominsky, A multi-institutional investigation of students' preinstructional ideas about cosmology, Astron. Educ. Rev. **11**, 46 (2012).

[7] F. Korur, Exploring seventh-grade students' and pre-service science teachers' misconceptions in astronomical concepts, Eurasia J. Math. Sci. Technol. Educ. **11**, 1041 (2015).

[8] A. Lelliott and M. Rollnick, Big ideas: A review of astronomy education research 1974-2008, Int. J. Sci. Educ. **32**, 1771 (2010).

[9] P. A'Hearn. Student understanding of scale and structure in astronomy: What classroom strategies are effective?, Master's thesis, California State University, San Bernardino, 2010, theses Digitization Project, 3863.

[10] M. Cole, C. Cohen, J. Wilhelm, and R. Lindell, Spatial thinking in astronomy education research, Phys. Rev. Phys. Educ. Res. **14**, 010139 (2018).

[11] M. Jones, G. Gardner, A. Taylor, E. Wiebe, and J. Forrester, Conceptualizing magnification and scale: The roles of spatial visualization and logical thinking, Res. Sci. Educ. **41**, 357 (2011).

[12] P. Sadler. The initial knowledge state of high school astronomy students, Ph.D. thesis, Harvard University, Cambridge, MA, 1992.

[13] R. Trumper, A cross-college age study of science and nonscience students' conceptions of basic astronomy concepts in preservice training for high-school teachers, J. Sci. Educ. Technol. **10**, 189 (2001).

[14] L. Shore and R. Kilburn, The effect of astronomy teaching experience on the astronomy interest and conceptions of elementary school teachers, in *Proceedings of the 3rd international seminar on Misconceptions and Educational Strategies in Science and Mathematics* (Misconceptions Trust, Ithaca, NY, 1993), Vol. 1.

[15] J. Mant and M. Summers, A survey of british primary school teachers' understanding of the earth's place in the universe, Educ. Res. **37**, 3 (1995).

[16] C. Bakas and T. Mikropoulos, Design of virtual environments for the comprehension of planetary phenomena based on students' ideas, Int. J. Sci. Educ. **25**, 949 (2003).

[17] W. Miller and F. Brewer, Misconceptions of astronomical distances, Int. J. Sci. Educ. **32**, 1549 (2010).

[18] V. Rajpaul, C. Lindstrøm, M. Engel, M. Brendehaug, and S. Allie, Phys. Rev. Phys. Educ. Res. **14**, 020108 (2018).

[19] J. Piaget, *The Child's Conception of the World* (Routledge & Kegan Paul, New York, 1929).

[20] U. Kanli, Using a two-tier test to analyse students' and teachers' alternative concepts in astronomy, Sci. Educ. Int. **26**, 148 (2015).

[21] P. Hewson and M. Hewson, Effect of instruction using students' prior knowledge and conceptual change strategies on science learning, J. Res. Sci. Teach. **20**, 731 (1983).

[22] R. Driver and J. Easley, Pupils and paradigms: A review of literature related to concept development in adolescent science students, Stud. Sci. Educ. **5**, 61 (1978).

[23] M. McCloskey, Naive theories of motion, in *Mental Models*, edited by D. Gentner and A. Stevens (Psychology Press, Hove, East Sussex, 1983), pp. 299–324.

[24] *Proceedings of the Second International Seminar: Misconceptions and Educational Strategies in Science and Mathematics*, edited by J. Novak (Cornell University, Department of Education, Ithaca, NY, 1987).

[25] N. F. Comins, *Heavenly Errors: Misconceptions about the Real Nature of the Universe* (Columbia University Press, New York, NY, 2001).

[26] S. Vosniadou and W. Brewer, Mental models of the earth: A study of conceptual change in childhood, Cogn. Psychol. **24**, 535 (1992).

[27] U. Kanli, A study on identifying the misconceptions of pre-service and in-service teachers about basic astronomy concepts, Eurasia J. Math. Sci. Technol. Educ. **10**, 471 (2014).

[28] P. M. Sadler, in *Proceedings of the Second International Seminar: Misconceptions and Educational Strategies in Science and Mathematics* (Cornell University, Ithaca, NY, 1987), Vol. 3, pp. 422–425.

[29] W. Bisard, R. Aron, M. Francek, and B. Nelson, Assessing selected physical science and earth science misconceptions of middle school through university preservice teachers, J. Coll. Sci. Teach. **24**, 38 (1994).

[30] M. LoPresto and S. Murrell, An astronomical misconceptions survey, J. Coll. Sci. Teach. **40**, 14 (2011).

[31] M. Zeilik and V. Morris, An examination of misconceptions in an astronomy course for science, mathematics, and engineering majors, Astron. Educ. Rev. **2**, 101 (2003).

[32] H. Bekaert, H. Van Winckel, W. Van Dooren, A. Steegen, and M. De Cock, Identifying students' mental models of the apparent motion of the sun and stars, Phys. Rev. Phys. Educ. Res. **18**, 010130 (2022).

[33] J. Plummer, A cross-age study of children's knowledge of apparent celestial motion, Int. J. Sci. Educ. **31**, 1571 (2009).

[34] S. Slater, S. Price Schleigh, and D. Stork, Analysis of individual test of astronomy standards (toast) item responses, J. Astron. Earth Sci. Educ. **2**, 89 (2015).

[35] H. Kalkan and K. Kiroglu, Science and nonscience students' ideas about basic astronomy concepts in preservice training for elemantary school teachers, Astron. Educ. Rev. **6**, 15 (2007).

[36] S. Serttaş and A. Türkoğlu, Diagnosing studentsâ misconceptions of astronomy through concept cartoons, Particip. Educ. Res. **7**, 164 (2020).

[37] R. Trumper, University students' conceptions of basic astronomy concepts, Phys. Educ. **35**, 9 (2000).

[38] A. Lightman and P. Sadler, Teacher predictions versus actual student gains, Phys. Teach. **31**, 162 (1993).

[39] P. Bitzenbauer, S. Navarrete, F. Hennig, M. Ubben, and J. Veith, Cross-age study on secondary school students' views of stars, Phys. Rev. Phys. Educ. Res. **19**, 020165 (2023).

[40] I. Testa, S. Leccia, and E. Puddu, Astronomy textbook images: Do they really help students?, Phys. Educ. **49**, 332 (2014).

[41] C. Chen, M.Schneps, and G. Sonnert, Order matters: Sequencing scale-realistic versus simplified models to improve science learning, J. Sci. Educ. Technol. **25**, 806 (2016).

[42] M. Cin, Alternative views of the solar system among Turkish students, Int. Rev. Educ. **53**, 39 (2007).

[43] L. Agan, Stellar ideas: Exploring students' understanding of stars, Astron. Educ. Rev. **3**, 77 (2004).

[44] E. Laski and R. Siegler, Is 27 a big number? correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison, Child Dev. **78**, 1723 (2007).

[45] L. J. Rips, How many is a zillion? sources of number distortion, J. Exp. Psychol. **39**, 1257 (2013).

[46] R. Siegler and J. Opfer, The development of numerical estimation: Evidence for multiple representations of numerical quantity, Psychol. Sci. **14**, 237 (2003).

[47] M. Ebersbach, A. Frick, K. Luwel, P. Onghena, and L. Verschaffel, The relationship between the shape of the mental number line and familiarity with numbers in 5- to 9-year old children: Evidence for a segmented linear model, J. Exp. Child Psychol. **99**, 1 (2008).

[48] E. Slusser, R. Santiago, and H. Barth, Developmental change in numerical estimation, J. Exp. Psychol. **142**, 193 (2013).

[49] A. Favia, N. Comins, G. Thorpe, and D. Batuski, A direct examination of college student misconceptions in astronomy: A new instrument, J. Rev. Astron. Educ. Outreach **1**, 21 (2014).

[50] K. Craik, *The Nature of Explanation* (University Press, Cambridge, United Kingdom, 1943).

[51] P. Johnson-Laird, *Mental Models Towards a Cognitive Science of Language, Inference, and Consciousness*, Cognitive Science Series Vol. 6 (Harvard University Press, Cambridge, MA, 1983

[52] J. K. Doyle and D. N. Ford, Mental models concepts for system dynamics research, Sys. Dyn. Rev. **14**, 3 (1998).

[53] M. Ubben, J. Hartmann, and A. Pusch, Holes in the atmosphere of the universe: An empirical qualitative study on mental models of students regarding black holes, Astron. Educ. J. **2** (2022).

[54] L. van Ments and J. Treur, Mental models and their dynamics, adaptation, and control: A self-modeling network modeling approach, in *Studies in Systems, Decision and Control*, 1st ed. (Springer International Publishing AG, Cham, 2022), Vol. 394, pp. 3–26.

[55] E. Corpuz and N. Rebello, Investigating students' mental models and knowledge construction of microscopic friction. i. implications for curriculum design and development, Phys. Rev. ST Phys. Educ. Res. **7**, 020102 (2011).

[56] J. Gilbert and C. Boulter, *International Handbook of Science Education*, edited by B. Fraser and K. Tobin (Kluwer Academic Publishers, London, 1998), pp. 53–56.

[57] H. Bekaert, Studying learning opportunities in a planetarium environment, Ph.D. thesis, KU Leuven, 2024.

[58] G. Özdemir and D. Clark, An overview of conceptual change theories, Eurasia J. Math. Sci. Technol. Educ. **3**, 351 (2007).

[59] D. Harlow and in Bianchini, *Science Education in Theory and Practice*, edited by B. Akpan and T. Kennedy, Springer Texts in Education (Springer International Publishing AG, Cham, 2020), pp. 389–401.

[60] A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).

[61] S. Vosniadou, The development of students' understanding of science, Front. Educ. **4**, 1 (2019).

[62] T. Konkle and A. Oliva, Canonical visual size for real-world objects, J. Exp. Psychol. **37**, 23 (2011).

[63] M. Szubielska and B. Bałaj, Mental size scaling of three-dimensional objects perceived visually or tactilely, Adv. Cognit. Psychol. **14**, 139 (2018).

[64] A. Gardony, M. Eddy, T. Brunyé, and H. Taylor, Cognitive strategies in the mental rotation task revealed by EEG spectral power, Brain Cognit. **118**, 1 (2017).

[65] R. Shepard and J. Metzler, Mental rotation of three-dimensional objects, Science **171**, 701 (1971).

[66] S. D. Muthukumaraswamy, B. W. Johnson, and J. P. Hamm, A high density ERP comparison of mental rotation and mental size transformation, Brain Cognit. **52**, 271 (2003).

[67] M. Fourtassi, G. Rode, and L. Pisella, Using eye movements to explore mental representations of space, Ann. Phys. Rehabil. Med. **60**, 160 (2017).

[68] R. Johansson, J. Holsanova, and K. Holmqvist, Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness, Cogn. Sci. **30**, 1053 (2006).

[69] M. Fourtassi, A. Hajjioui, C. Urquizar, Y. Rossetti, G. Rode, and L. Pisella, Iterative fragmentation of cognitive maps in a visual imagery task, PLoS One **8**, e68560 (2013).

[70] T. Tretter, M. Jones, T. Andre, A. Negishi, and J. Minogue, Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena, J. Res. Sci. Teach. **43**, 282 (2006).

[71] S. Jung, S. Roesch, E. Klein, T. Dackermann, J. Heller, and K. Moeller, The strategy matters: Bounded and unbounded number line estimation in secondary school children, Cognit. Dev. **53**, 100839 (2019).

[72] J. Sharp, Children's astronomical beliefs: A preliminary study of year 6 children in south-west England, Int. J. Sci. Educ. **18**, 685 (1996).

[73] T. Makwela, Probing student engagement with size and distance in introductory astronomy, Ph.D. thesis, University of Cape Town, 2022.

[74] C. Eames, R. Eames, and K. Boeke, *Powers of Ten* (1978).

[75] K. Yu, K. Sahami, and J. Dove, Learning about the scale of the solar system using digital planetarium visualizations, Am. J. Phys. **85**, 550 (2017).

[76] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman, The science of visual data communication: What works, Psychol. Sci. Publ. Interest **22**, 110 (2021).

[77] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini, How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques, in *CHI 2015: Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems: Seoul, Republic of Korea* (ACM, New York, NY, 2015), Vol. 2015, pp. 1469–1478.

[78] A. diSessa, Knowledge in pieces, in *Constructivism in the Computer Age*, edited by G. Forman and P. Pufall (Lawrence Erlbaum Associates, Hillsdale, NY, 1988), pp. 49–70.