# Extracting information from the Hertzsprung-Russell diagram: An eye-tracking study

Ronja Langendorf,* Susanne Schneider, and Pascal Klein
*Faculty of Physics, Physics Education Research, Georg-August-Universität Göttingen, Friedrich-Hund-Platz 1, 37077 Göttingen, Germany*

(Received 30 April 2022; accepted 18 August 2022; published 27 September 2022)

The Hertzsprung-Russell diagram (HRD) is a fundamental representation in stellar physics. It contains information about key properties of stars and allows inferences about stellar evolution. The use of the HRD is an important disciplinary activity in astrophysics. For example, it is particularly important to have a graphical understanding of the HRD in order to understand elementary astrophysical relationships (e.g., about the luminosity, temperature, radius, and mass of stars). However, several research papers indicate that students often have difficulty interpreting the HRD, apparently due to its visual complexity, and a number of learning difficulties have been described. Yet, there is still no evidence concerning how learners actually select and extract information from the HRD when completing tasks. In this study, we examined the gaze patterns and think-aloud protocols of 35 physics students as they performed 14 open-response tasks. Benchmarking against traditional *x*-*y* diagrams shows that the HRD imposes a significantly higher cognitive load on students, particularly due to the representation of luminosity, magnitude, and spectral class. Students reported a variety of learning difficulties related to information selection and extraction, sometimes mechanistically copying procedures from typical *x*-*y* diagrams. Eye-movement analysis confirmed these learning difficulties on a procedural level and show whether the students fixated on task-relevant parts of the HRD. Based on the study results, preliminary recommendations can be made in order to create engaging learning materials relating to the HRD.

DOI: 10.1103/PhysRevPhysEducRes.18.020121

## I. INTRODUCTION

In astrophysics, the Hertzsprung-Russell diagram (HRD) is considered the most common representation for plotting stellar quantities. A typical HRD like the one in Fig. 1 essentially contains information about the brightness and temperature of stars, but astrophysicists extract much more from it. Disciplinary knowledge can be used to draw further conclusions about quantities and properties that are not explicitly plotted and that are related, for example, to the stages of stellar evolution [1]. The diagram has historical roots that are reflected in this representation, but it is still used as an important tool in stellar physics and in university teaching [2]. The HRD is said to have several difficulties that can be challenging, especially for new learners interpreting the diagram [1,3]. Unfortunately, there is a gap in empirical studies addressing this issue and investigating the use of this important diagram with such unique properties. In physics education research (PER), learners' use of diagrams is a current topic. One focus is the use of diagrams in mechanics and kinematics, where a number of tests have been developed and used to examine learners' content knowledge and understanding [4–7]. In addition, interpreting diagrams requires highly visual and cognitive processes, and to better understand these, eye tracking (ET) has been used in recent research projects [8–11] (for a systematic review, see Ref. [12]). ET can be used to study the problem-solving process when dealing with representations such as the HRD. In this study, we will investigate the difficulties in interpreting the HRD from the student perspective, paying particular attention to cognitive load (CL) and visual attention when working with the HRD. To this end, we first discuss the properties of the HRD and the associated learning difficulties that motivate the research questions. Building on previous research on diagrams that involves eye tracking, the method and material of this study are explained. This is followed by the presentation of the results and a concluding discussion.

## II. THEORETICAL BACKGROUND

### A. Origin and physical properties of the Hertzsprung-Russell diagram

In the 19th century, the photography of stellar spectra increasingly became a focus of research interest. At the time this was the only way to obtain information about the

*rlangen@uni-goettingen.de

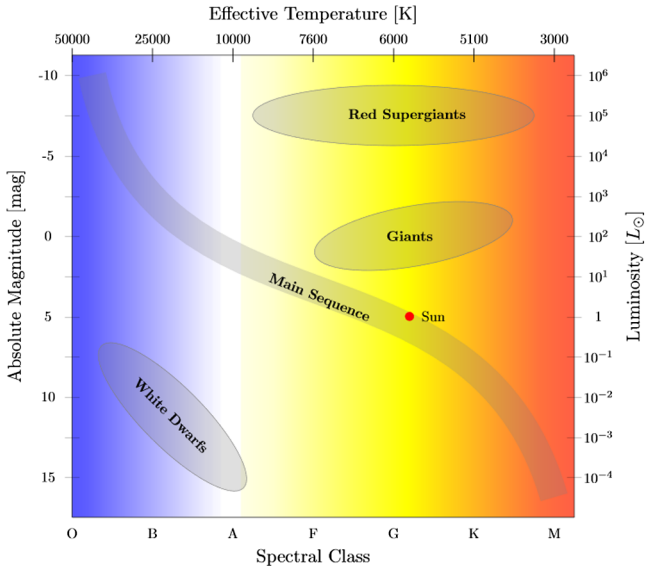


FIG. 1. A typical Hertzsprung-Russell diagram that contains information about central stellar properties. The spectral class and effective temperature (in kelvins) are shown in the horizontal direction, and the absolute magnitude (in mag) and the luminosity (in solar luminosity) are shown in the vertical direction. The colored background represents the color of the star. This example contains the Sun and four luminosity classes that are part of the stages of stellar evolution (see text for details).

temperature of these celestial objects. The American physician and astronomer Henry Draper therefore pioneered the science of photographing and analyzing these spectra [13]. He succeeded for the first time in recording the lines of a stellar spectrum, and his work led to the so-called Draper catalog of the spectra of stars. In it, stars with different brightnesses are classified alphabetically (A, B, C,...) based on their spectra and the wavelengths of their lines [14]. Annie Jump Cannon and Edward Pickering were part of a research team that refined Draper's work and, based on new findings, developed the so-called Harvard classification of stars, that is still used today. Taking into account temperature and luminosity, stars have since been classified into spectral classes O, B, A, F, G, K, and M [15,16]. The relationships between spectral class, temperature, luminosity, and brightness were then echoed in the well-known HRD. This diagram can be traced back to the astronomers Ejnar Hertzsprung (1873–1967) and Henry Norris Russell (1877–1957) [15], who independently published earlier versions of it [17,18]. In a classical and original HRD, absolute visual magnitude as a measure of brightness is plotted as a function of spectral class [18]. The typical HRD as used today also includes information on effective temperature, luminosity, and star color (Fig. 1). The $x$ and $y$ axes of the HRD are therefore defined twice. As mentioned before, the spectral class is plotted on the classical abscissa. It is a historically based measure of the temperature sequences located on the upper abscissa. The effective temperature $T_{\mathrm{eff}}$ (given in kelvins [K]) increases on a nonlinear scale from right to left and can be represented approximately by the star color (background of the diagram). The temperature is related to the luminosity $L$ of a star, which is described by the Stefan-Boltzmann law given by [15]

$$L = 4\pi R^2 \sigma T_{\mathrm{eff}}^4, \tag{1}$$

where $\sigma = 5.6704 \times 10^{-8}\ \mathrm{Js^{-1}\,m^{-2}\,K^{-4}}$ is the Stefan-Boltzmann constant and $R$ is the radius of the star. The luminosity is plotted with a logarithmic scale on one of the HRD's $y$ axes and is normalized to the solar luminosity $L_\odot$. Equation (1) shows that the radius of stars plotted in the HRD increases from the lower left to the upper right. To illustrate this relationship, there are also some HRD examples with parallel lines indicating a constant radius. On the classical ordinate, the brightness is plotted as *absolute* magnitude $M$, which is defined as the *apparent* magnitude $m$ of a star at a standard distance of 10 parsecs. It is given by [19]

$$M = M_\odot - 2.5\ \log\ L, \tag{2}$$

where $M_\odot$ is the absolute magnitude of the Sun. According to Eq. (2), a star with a 100-fold luminosity differs from another star by 5 magnitudes. This definition of magnitude also has historical roots, dating back to the approach of designating the brightest star visible to the naked eye as magnitude 1 and the star just visible to the naked eye as magnitude 6. As telescopes advanced, stars even dimmer than 6 magnitudes became visible [20]. The historical scale was also extended to include negative magnitude values for stars even brighter than 1 magnitude. Consequently, the brightest stars with the smallest magnitudes are at the top of the HRD.

An important finding was that stars are not evenly distributed in the HRD but cluster in certain areas. Accordingly, the stars have a set of fixed physical properties they are assigned. Four of the so-called luminosity classes are shown in Fig. 1—main sequence, giants, red supergiants, and white dwarfs. Today, we know that these four luminosity classes represent the stages of stellar evolution. The life of a star begins with the main sequence, and the more massive it is, the higher its temperature and luminosity (*mass-luminosity relationship*). Main sequence stars burn hydrogen to luminate, and the more massive they are, the faster they do so. Therefore, their lifetime is about $10^3$ years shorter than that of the Sun ($10^{10}$ yr), and they leave the main sequence first. Therefore, by plotting star clusters on the HRD, one can infer the age of the stars from the structure of the main sequence. In short, the subsequent burning phases (helium, carbon, etc.) are shorter and hotter. The star expands and transforms from a giant to a red supergiant. After a large energy loss, low-mass stars are left with only their hot cores, and they become white dwarfs,

denoting the end of star life [15]. (Massive stars instead become neutron stars or black holes, neither of which can be represented in an HRD).

### B. Learning difficulties associated with the Hertzsprung-Russell diagram

It is clear that these four stages of stellar evolution are the result of complex physical processes that cannot be discussed in detail here. It is also obvious that the HRD contains a large amount of information and relationships pertaining to stars, and for astrophysicists it is much more than just a visualization of physical values. This aspect of the HRD is described by Airey and Eriksson as high *disciplinary affordance* [1,21]. Accordingly, astrophysicists are disciplinary insiders who can automatically read the visible and invisible information of the HRD [1]. In contrast, they consider the diagram to be unlikely to convey this content in an educational context and therefore evaluate its *pedagogical affordance* [21] as low. This rating is based on several learning difficulties that may become relevant for novice learners as opposed to astrophysicists. Learning difficulties may arise because (i) the definitions of absolute magnitude and spectral classes have their roots in *history* and may seem atypical today, (ii) there is an *omission* of central stellar quantities (radius, mass), (iii) the HRD has an *overloading* nature due to the amount of information, and (iv) some aspects are counterintuitive and do not meet the usual *expectations* of students. The latter refers, for example, to the expectations that scales increase linearly from left to right or that a hot temperature is automatically associated with the color red. This color-temperature relationship refers to the classification of colors as it is done within color theory. Hereby, Airey and Eriksson follow diSessa's understanding of phenomenological primitives (*p* prims [22]) when assuming that learners use these everyday heuristics [1]. Moreover, in everyday heuristics, deadness is associated with coldness, but white dwarfs are particularly hot. Table I lists these potential barriers and students' possible learning difficulties in dealing with the HRD. Building on their analysis, Airey and Eriksson conclude that students are therefore unlikely to be able to recognize at first glance the disciplinary affordance that an astrophysicist identifies in the HRD [1].

In general, the ability of learners to use diagrams for problem solving is essential and has already been studied in depth in contexts very different from astrophysics (especially in mechanics; see Sec. II C). Regarding the HRD, there is very little empirical research on it, although it is very important and very useful in the context of astrophysics. This diagram is a very common learning tool in introductory astronomy courses at universities around the world (especially in the US [2]), and it is also covered in mainstream pedagogical literature, such as that of Carroll and Ostlie [23].

### C. Diagrams and eye tracking

Understanding graphs is a necessary prerequisite for learning in most higher education subjects [24], as graphs can be used to simplify abstract concepts and facilitate the exchange of information between individuals [25,26]. Apart from this, the competence to work with graphs is a key aspect of general skills, such as media literacy [27], online reasoning [28], data literacy [29], and information problem solving [30]. In PER, there is extensive research on the difficulties learners have in dealing with graphs [4–7]. Early work indicates that students confuse slope and height on a graph, have difficulty interpreting changes in height and slope, and have problems with the concept of area under a curve [31]. Based on significant research on students' difficulties with the graphical representation of position, velocity, and acceleration versus time, the Test of Understanding Graphs in Kinematics (TUG-K) was developed [4]. Since then, the TUG-K has been used in PER to evaluate students' understanding of kinematics, and it has served as a reference for developing further tests. However, research on graphs is not limited to the fields of kinematics and mechanics. Dealing with graphs in general involves highly visual processes, such as selecting information, scanning the coordinate system to locate a target point, reading axis information (such as orientation), and integrating information between the graph and the accompanying text. Beyond the mere outcome of the assessment (i.e., correct or incorrect responses), the problem-solving process contains information about thinking patterns, solution strategies, and task characteristics, and recent PER studies have shed light on the problem-solving process itself. To

TABLE I. Potential learning difficulties expected for students extracting information from the HRD (Fig. 1) based on Airey and Eriksson [1].

| Potential barriers | Possible learning difficulties with the HRD |
|---|---|
| History | Magnitude: Scale definition and interpretation; spectral class: Scale definition and interpretation |
| Omission | Radius: Increase from bottom left to top right; mass (main sequence): Increase with increasing luminosity |
| Overloading | Colored background: Not for aesthetic reasons; visual complexity: Wealth of information |
| Expectations | Temperature-axis: Hotter is to the left; color–temperature relationship: Hot is blue, cold is red; life–death analogy: a "dead" star is hot; logarithmic scale: temperature and luminosity scales are not linear |

this end, visual attention while working with diagrams has been captured using eye tracking [8–11]. As working with the HRD imposes high demands in terms of the reception and extraction of information due to its inherent complexity, it is worth taking a closer look at these processes at a visual-perceptual level.

In the following, we briefly summarize the most relevant work from PER and peripheral fields that investigated information retrieval from diagrams using eye tracking. Chumachemko *et al.* studied the eye movements of novices and experts while navigating to points in a Cartesian coordinate system [32]. They found that both groups performed saccades in vertical and horizontal directions more frequently than in other directions, which they interpreted as evidence of "theoretical" perceptual actions. These saccades reflect the cultural way of approaching the Cartesian coordinates system. Klein *et al.* found a similar result when they examined physics students' viewing of vector field diagrams [33,34]. The vector field diagrams were displayed in a two-dimensional Cartesian coordinate system, and students were asked to judge whether the divergence of the vector fields was zero or nonzero. Klein *et al.* found that students who performed predominantly horizontal and vertical saccades were more likely to obtain a correct result. Eye movements in the horizontal and vertical directions indicated that students were comparing adjacent vectors; that is, they were applying a rigorous procedure to interpret and determine divergence [33,34]. In another study investigating students' eye movements when dealing with line graphs, Klein *et al.* reported that physics students traced the line graphs more frequently with their eyes than nonphysics students. That is, the physics students performed saccadic eye movements that corresponded to the gradient angle of the line graph [10]. In particular, when qualitatively comparing the slope of two graphs, most eye movements followed the graph. This was interpreted as correct cognitive processing of the slope concept or in the words of Chumachemko *et al.* as evidence of "theoretical" perceptual actions.

In the above-mentioned study by Susac *et al.* [8] and its replication by Klein *et al.* [10], unknown axis labels were consistently reported to receive more attention. Both studies used isomorphic pairs of tasks (i.e., one set of tasks was framed in a physics context, while the other was framed in an everyday finance context) that were presented to physics and nonphysics students. The physics students spent more time on the finance axis labels and the nonphysics students spent more time on the physics axis labels. Thus, both groups of students needed more time to extract information from the axes of graphs the context of which was unfamiliar to them.

Overall, the above-mentioned studies showed that eye tracking is an appropriate method to investigate students' cognitive processes when using diagrams. In summary, the reported results suggest some implications that are important for the study of visual interaction with the HRD. In particular, the studies considered provide information on the areas of interest (AOIs) and the eye-tracking metrics that are important in investigating the HRD.

### D. Cognitive load, expertise, and information reduction

*Cognitive load theory* (CLT) is concerned with the role of CL *vis-à-vis* working memory in learning and problem solving [35]. The theory assumes that CL and limited working memory capacity affect learning outcomes. A distinction is made between *intrinsic cognitive load* (ICL), *extraneous cognitive load* (ECL), and *germane cognitive load* (GCL) [36]. First, ICL is caused by the complexity of learning content and information itself and can only be modified if either the learner's prior knowledge or the learning content change [37]. It is determined by the level of *element interactivity* [36,38,39]. Here, elements means "any information that needs to be learned" [[38] p. 305]. When elements can be learned independently and sequentially and have few connections to other elements (i.e., isolated memorization), element interactivity and ICL are low. In contrast, deep understanding of interacting elements is associated with high element interactivity and, consequently, high ICL [36]. At this point, the learner's prior knowledge becomes relevant. The grouping of several elements that are strongly related (and distinct from other element groups) into a single unit is called *chunking*; correspondingly, this unit is called a chunk [40]. In addition to the often-cited example of chess [41], chunking effects also occur in problem solving and learning with multiple representations in physics [42]. Similar to the CLT's assumption, it is assumed here that only a limited number of chunks can be cognitively processed simultaneously [43]. Consequently, it can be said that learners with more prior knowledge and more expertise are better able to extract information from learning materials by forming chunks [40], something that cognitive unloading is associated with [44]. This ability can also be explained by the *information-reduction hypothesis* [45,46]. According to this hypothesis, learners with expertise select relevant information and focus their attention on it. Thus, cognitive processing is limited to task-relevant information and ideally ignores task-redundant information. For the HRD, the related elements such as color, temperature, and spectral class could be grouped in order to reduce cognitive resources.

Besides the intrinsic complexity of learning materials, external factors such as instructional aspects or presentation of learning material can put an ECL on learners. Unlike ICL, which is relevant for learning, the cognitive processes related to ECL are not productive for learning [47]. To better cope with ICL, the material should be designed so that ECL is not unnecessarily increased [35,36,44]. Last, GCL is related to working memory resources and caused by the learning process itself. It must be applied to deal with ICL to enable learning [36]. As a result of discussing the

reasonable number of CL types [48], GCL, as originally characterized in 1998, is redefined and currently no longer seen as a separate CL type. Sweller *et al.* describe that GCL redistributes cognitive resources from extraneous to intrinsic aspects and does not generate its own load [37]. This redistribution in favor of increased capacity for ICL is also called *germane processing* [47]. Consequently, GCL or rather germane processing can be considered positive, as it is productive for learning.

In this study of the HRD, the interactivity of the elements in the context of ICL plays the most important role. The HRD can be considered as a representation with high element interactivity, since the information contained on the stars cannot be considered in isolation and requires deep understanding. Therefore, the simultaneous inclusion of the different elements of the HRD can be difficult. The concept of ICL can help explain learners' difficulties with the HRD, although element interactivity alone is not sufficient to assess the difficulty of the learning material [48].

## III. RESEARCH QUESTIONS

In this study, we investigate the potential of eye tracking when first-year students extract information from the HRD. Our goal is to explore students' difficulties and mastery of the HRD (both on an outcome and on a process level), as well as strategies used by the participants. Eye-tracking data will be supplemented by performance measurement results and results from postinterviews. The following research questions will be addressed in this study

(1) How well do the students succeed in extracting information from the HRD (as measured by test scores), and what difficulties do they report (as revealed by student interviews and CL ratings)?
(2) Can the reported difficulties working with the HRD be empirically supported by the analysis of visual attention?
(3) How does the level of expertise impact processing of the HRD?

The first research question relates to students' learning difficulties with the HRD, which primarily become evident at the outcome level. Here, we aim to corroborate and extend the findings of previous studies on the understanding of the HRD. Additionally, we are particularly interested in the pieces of information the students allocate their attention to. Therefore, according to the second research question, we investigate learners' visual attention while they perform items, including the HRD. To quantitatively measure the difficulties in processing the HRD, we benchmark against analogous items using common *x*-*y* diagrams. By comparing the visual attention between the two types of diagrams for very similar item requirements, the difficulties can be quantified at a process level.

The last research question investigates how students with a high outcome level handle the HRD compared to those with a low one to solve the items. From the gaze data of the high levels compared to the low levels, we can learn what dealing with the HRD looks like when it succeeds. In conjunction with the collected thought processes, the differences in the gaze behavior associated with high and low levels will help us to understand how both groups approach the HRD. Subsequently, a set of heuristics (i.e., approaches to problem solving) could be determined that are also relevant to the second research question.

## IV. METHOD

### A. Data collection and sample

The study took place during the winter term of 2021–22 at the University of Goettingen. Thirty-five freshmen physics students took part in the study. The demographics of the participants are summarized in Table II. We recruited the students during lecture time without explicitly telling them what the study was about. Participation was voluntary, and the students were compensated with 15 euros. We assumed that the participants were basically able to extract information from graphs. We chose first-semester physics students because we can assume that cognitive errors, such as interpreting graphical data as an image, are not represented (and our results confirm this). In the eye-tracking lab, we asked students if they had seen or worked with the HRD before. Only one student was familiar with the HRD. All participants had normal or correct-to-normal vision.

### B. Study design and procedure

The experiment consisted of several parts in three superordinate phases, which are shown in Table III. All participants came one by one to our eye-tracking lab, where a researcher (R.L.) guided them through the entire experiment, so that the students were not alone at any point. Meanwhile, the researcher stayed in the background, guided through, and supervised the experiment to ensure the quality of the data. After the students gave informed consent to participate in the experiment, they were seated in front of a 24-inch computer screen fitted with a stationary eye-tracking system (Tobii X3-120, <0.40° spatial accuracy, 120 Hz sampling rate, about 65 cm distance). A 9-point calibration procedure was used to record eye movements, and the students were instructed not to move their heads if

TABLE II. Summary of students' demographics (mean ± standard deviation).

| | Students |
|---|---|
| Sample size | 35 |
| Female | 10 |
| Age | 19.7 (1.5) |
| Grade of high school diploma[a] | 1.5 (0.5) |
| Spatial abilities[b] | 0.61 (0.17) |

[a]Ranging from 1.0 (best) to 4.0 (German average: 2.37 [49]).
[b]Ranging from 0.0 to 1.0 (best) (*high* as of 0.52 [50]).

TABLE III. Elements of the experimental procedure [eye tracking (ET)].

| Phase | description | Duration |
|---|---|---|
| 1a | Participants take five context free-diagram (CFD) items (ET); subsequently, they take a CL questionnaire | 10 min |
| 1b | Participants provide information about their prior experience with the HRD and with astrophysics | 3 min |
| 2a | Participants solve 14 HRD items (ET); subsequently, they take a CL questionnaire and judge the item difficulties on a rating scale | 25 min |
| 2b | Guided (retrospective) student interviews | 15 min |
| 3 | Demographics and spatial span task | 7 min |

possible. In the first phase of the experiment (phase 1a), the participants answered five items related to the location of points in a Cartesian ($x$-$y$) coordinate system—a context-free diagram (CFD). The requirements in these items correspond to the requirements for the HRD later in the experiment. Figure 2 shows one such CFD item (item 3 of 5) and the corresponding HRD item (item 6 of 14), and Table IV gives an overview of all items, with isomorphic pairs indicated by an asterisk. The exact procedure of the stimuli presentation and response process is depicted in Fig. 3 and is described in detail below. After answering the five CFD items, the participants completed a short CL questionnaire asking them about the CL caused by the items they had just completed.

Before beginning the second phase of the experiment, the students were asked about their prior experience with astrophysics (in school or in their spare time) and whether they already knew anything about the HRD (phase 1b).

At the beginning of the second phase of the experiment, calibration of the eye-tracking system was performed again.

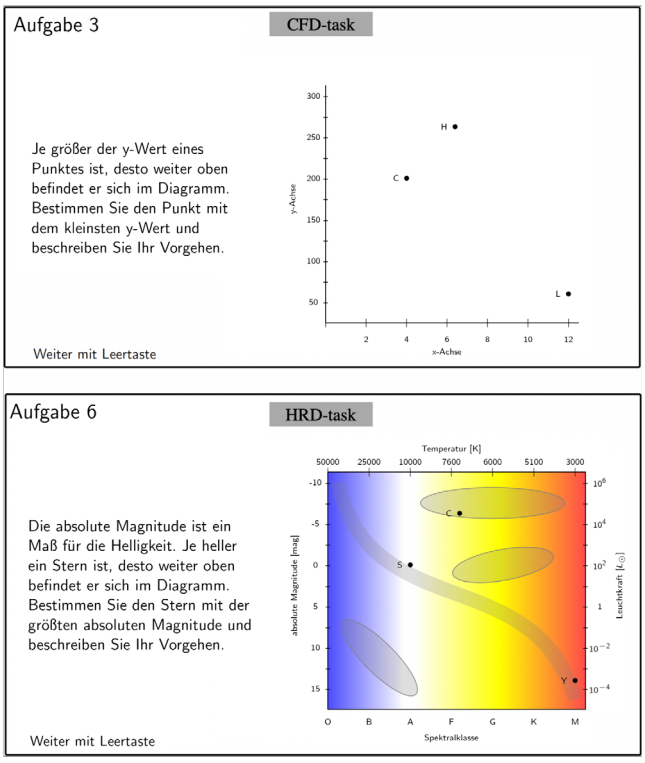


FIG. 2. Isomorphic pair of items as originally used in the study (English translations can be found in the Supplemental Material [51]). The top item shows three points in an $x$-$y$ coordinate system (C,H,L), and students must compare the $y$ values to find the point with the largest $y$ value. The bottom item shows three stars (S,C,Y), and students must select the star with the largest absolute magnitude. Note the parallel item design, especially the congruent position of the points or stars in the diagrams.

TABLE IV. A description of the 14 items related to the HRD. Analogous items with common $x$-$y$ diagrams were used in addition (indicated by the † symbol).

| Item | Item requirement | Given information |
|---|---|---|
| T1† | Extract the spectral class<br>(† extract the $x$ value) | 2 stars<br>(†2 points) |
| T2† | Predict a development for increasing $T$ while $L = \text{const}$<br>(† predict a development for increasing $x$ while $y = \text{const}$) | 1 star<br>(†1 point) |
| T3 | Compare $T$ values depending on spectral class | 3 stars |
| T4 | Predict a development of color depending on $T$ | 1 star |
| T5 | Compare $T$ values depending on color | Giants, White dwarfs |
| T6† | Compare $M$ values<br>(† compare $y$ values) | 3 stars<br>(†3 points) |
| T7† | Compare $L$ values<br>(† compare $y$ values) | 2 stars, Sun<br>(†2 points, reference point) |
| T8† | Compare pairs of $L$ values<br>(† compare pairs of $y$ values) | 3 pairs of stars<br>(† 3 pairs of points) |
| T9 | Discern a $M - L$ relation [Eq. (2)] | 2 stars |
| T10 | Discern a relation for main sequence stars concerning $T$, $L$, $M$ and color | Main sequence |
| T11 | Discern a $m$-$T$ relation by $L \propto m$ | Main sequence |
| T12 | Describe an $R$ development by $L \propto T \cdot R$ | Stellar development (3 phases) |
| T13 | Describe a stellar development in total plus radius | Stellar development (4 phases) |
| T14 | Compare $R$ values and color | 4 luminosity classes |

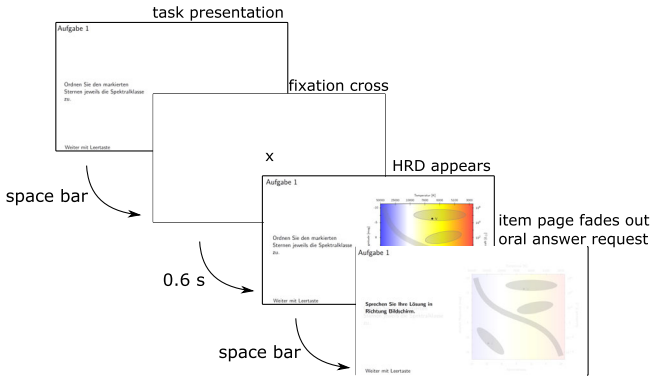


FIG. 3. Procedure of item presentation and answer submission as originally used in the study (English translations can be found in the Supplemental Material [51]). The item text appears on the left half of the screen. The corresponding item diagram is displayed under the user's control. In between, the user looks at a fixation cross. When the participant has solved the item in his or her head, he or she presses the space bar and gets to the input page. Here, the background fades out, and the prompt to speak the answer appears.

Next, the participants were presented with 14 items on the HRD (see Sec. IV C 1 for details on the items). The procedure for the items was exactly the same as for the previous items and is shown in Fig. 3. To begin, the students were shown the question statement without the HRD. By pressing the space bar, the students indicated that they read the question. Then, a white screen appeared with a fixation cross to set a starting point for the eye movements. This page automatically disappeared after 0.6 s, and the full item page (including the question statement and the HRD) appeared. The participants solved the item in their head and indicated that they were ready to give an answer by pressing the space bar again. In the final step, the participants spoke their answer aloud and the verbal responses were recorded via the computer's built-in microphone. During this procedure, the item was displayed with low contrast. This procedure was practiced with the students at the beginning of phase 1a using some trivial examples until the sequence of actions was understood. The rationale behind this sequence was twofold. First, we captured students' eye movements as they interacted with the HRD without the cognitive burden of parallel speaking and thinking after reading the question. Nevertheless, eye movements were recorded as the students spoke during the last presentation slide of each item. Second, we established an order whereby the question was read first and then the diagram was viewed, with a unique starting point for every student and item (using the fixation cross). In this way, transitions back to the question could be traced very easily. The slide on which the HRD appears is the main resource for analyzing eye-movement data (see Sec. IV D).

After completion of all items, the participants answered a questionnaire with nine items on the CL caused by the items they had just completed. In addition, all 14 HRD items were presented again on paper, and the test takers had to indicate the *perceived* difficulty of each item individually on a 6-point Likert-type rating scale.

After completion of this phase (2a, cf. Table III), the students were interviewed. The interview protocol included questions about difficulties the participants encountered while working with the HRD. In addition, the students were asked about individual strategies for using the diagram. Last, the participants provided demographic information and completed a standardized spatial span task on the computer to assess their spatial abilities.

### C. Materials

All study material is included in the Supplemental Material (translated from the original German into English), consisting of the HRD items and the CFD items presented during the eye-tracking study, short rating-scale questionnaires, and the interview protocol [51].

#### *1. HRD items and isomorphic CFD items*

In this study, students completed 14 HRD items and five isomorphic items with context-free *x*-*y* diagrams. The items require extracting and comparing values, describing and predicting developments, and recognizing relationships (see Table IV). The characteristics of the HRD are the double-sided axes, the logarithmic scaling, and the counterintuitive direction of the axis. We assume that the difficulty in extracting information from the HRD is due to these features, and we have designed the items in such a way that lack of content knowledge about astrophysics does not affect item elaboration. Therefore, astrophysics knowledge is of minor importance in answering the questions, as in this study the extraction of information from the HRD is required for scientific reasoning. As we do not have access to a standardized test from the literature that meets our requirements, we developed an instrument ourselves. In doing so, we followed the three steps of Airey and Eriksson as a recommendation for teachers when implementing the HRD [1]: The item demands increase as the experiment progresses. First, the students engage with the variables that are explicitly presented in the HRD (items 1–8); second, they are asked to identify the central relationships between the axes (items 9–12); and third, they must describe the key meaning of the diagram for astrophysics (items 13–14). This order was chosen to increase the pedagogical affordance and to help learners better understand the disciplinary utility of the HRD [1].

#### *2. Questionnaires*

Four questionnaires were used in this study. The CL scale was adopted from the literature [52] in line with the learning materials used in this study (i.e., the *x*-*y* diagrams and the HRD). The three-factor structure of the instrument

was analyzed (Kaiser's criteria and scree plot) and reflects the three CL types. Because of floor effects, the items related to ECL with respect to the $x$-$y$ diagram could not be included in the data analysis. These floor effects are also manifested in the low reliability of this factor ($\alpha = 0.08$). The questionnaires assessing *perceived* item difficulty ($\alpha = 0.72$, 6-point Likert-type rating scale) and familiarity with the HRD ($\alpha = 0.83$, 10-point Likert-type rating scale) were newly developed for this study and showed adequate reliabilities. For the assessment of spatial ability, we used a version of the spatial span task (SST) by Ref. [53]. The SST measures the ability to simultaneously process and hold spatial information in memory. The subjects must judge and recall the spatial representation of letter sets presented on the screen. A letter (F, J, L, P, or R) is (i) presented either correctly or mirrored and (ii) rotated in the presentation plane. Within 2 sec, the subjects must determine whether the representation of the letter is correct or mirrored by pressing the appropriate key, memorizing the spatial orientation of the letter, and recalling it for varying test set sizes.

### 3. Student interview protocol

During the guided student interviews, participants were given a paper representation of the HRD to support their retrospective reporting. The following guiding questions were used to encourage students to freely and openly discuss the previous phases 1a and 2a of the study: (i) How did you experience the item processing in this study? (ii) To what extent did you feel it was easy or difficult to deal with the HRD? Following the students' free reporting, more detailed follow-up questions were asked. On the one hand (process related), these referred to the specific actions during the processing and to the strategies used for dealing with the diagram (e.g., How did you PROCESS the items?). On the other hand (object-related), it was about the students' first impression of the HRD and its special features and characteristics (e.g., Can you describe the situation in which you saw the HRD for the first time?). The entire interview protocol can be found in the Supplemental Material [51].

## D. Data analysis

The spoken answers were evaluated according to a scheme that can be found in the Supplemental Material [51]. Correct responses to questions were scored with one point each. Depending on the items, between 1 (e.g., T11) and 4 points (e.g., T12) could be achieved. The ratio of the achieved score to the possible total score defines the difficulty (or solution probability) of an item. Two independent raters scored the answers and reached consensus ($\kappa = 0.91$).

Scores on the rating scales were linearly transformed to the interval [0,1], where 0 indicates low prior experience with stars or star development or low CL. Note that 0 indicates high item difficulty or high *perceived* item difficulty. Thus, the difficulty is reversed.

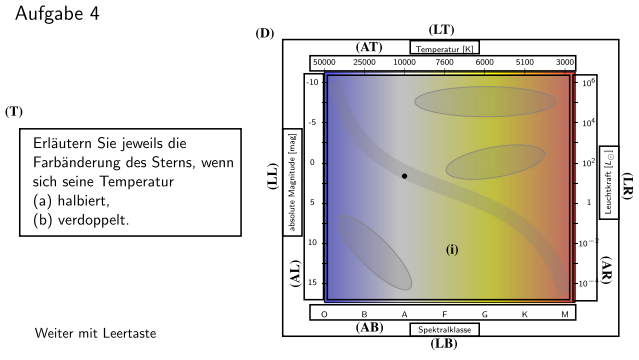


FIG. 4. Definition of areas of interest (AOIs) for all items as originally used in the study (English translations can be found in the Supplemental Material [51]). Two global AOIs cover the *text* (T) and the *diagram* (D). The D-AOI includes nine smaller, local AOIs, namely, the interior of the diagram (i) and each axis (AL, AR, AT, AB) and axis label (LL, LR, LT, LB) in the left, right, top, and bottom.

As shown in Fig. 4, two AOIs, *text* (T) and *diagram* (D), are defined for each item. The text contains the question and, if necessary, additional information, such as an explanation or an equation. The second AOI (D) covers the entire diagram, including the labeling of the axes. Its size does not differ between items, unlike the T-AOIs. In addition, the D-AOI consists of nine smaller AOIs that are the same for all items. These do not overlap, but have different sizes. The interior of the diagram is defined as i-AOI. Each axis (A) and each axis label (L) are defined as a single separate AOI for all four plotted variables (top (T), bottom (B), and right (R), left (L)), giving a total of nine local AOIs. According to Hahn and Klein, this definition of AOIs can be considered as both *global* (T, D) and *local* AOIs (AL, AR, AT, AB, i, LL, LR, LT, LB) [12]. The HRD and CFD item pairs are compared using paired $t$ tests with a Bonferroni correction. Here, the eye-tracking metrics total visit duration (TVD) and total visit counts (TVCs) are analyzed. The TVD provides the time duration for information access on an AOI as the sum of all fixation and saccade durations. A high value indicates higher visual attention on the AOI [12]. The TVCs correspond to the number of jumps into an AOI and indicate, for example, the rereading of a text when additional information from an image is integrated [12].

The audio recordings of the interviews were transcribed and analyzed using Mayring's qualitative content analysis [54]. The transcripts can be provided upon request. Using an inductive approach, all transcripts were analyzed in terms of difficulties, confusion, and problems related to the HRD expressed by the students. The resulting category system includes 10 main categories (see Sec. V D), and the coding of students' statements according to that system works well, as indicated by a good interrater reliability

(Cohen's kappa $\kappa = 0.78$). After discussing the independent ratings, the reliability was increased to $\kappa = 0.86$.

## V. RESULTS

In the following, the results of the study are presented by first addressing the prerequisites of the students, such as their specific prior knowledge in astrophysics. Then, the students' performance in problem solving is reported by analyzing and comparing, among other things, the item difficulty for both isomorphic phases (1a, 2a). In addition, perceived item difficulty and CL are analyzed. This is followed by an item-level analysis of the eye-tracking data, comparing isomorphic pairs and comparing high- and low-scoring students. The chapter concludes with interview results reporting learning difficulties and learning strategies.

### A. Prior knowledge and prerequisites

The mean of the participants' high school graduation score is 1.50 (0.50), which is better than the German average of $M = 2.37$ ($SD = 0.12$) in the 2019–2020 school year [49]. Students' prior experience with the HRD and stellar evolution was assessed with six rating-scale items ($\alpha = 0.83$), ranging from 0.00 (no prior knowledge) to 1.00 (high prior knowledge). This scale includes statements about familiarity with the topic. On average, students had low familiarity with this content ($M = 0.2$, $SD = 0.06$). In addition, about a third of the participants (12 out of 35) had not taken astrophysics courses, either in formal or informal learning environments. The analysis of the SST by [53] to assess spatial ability yielded a mean score of $M = 0.61$ ($SD = 0.17$). Thus, according to [50], students have a high spatial ability on average (SST $> 0.52$). Overall, the subjects can be considered high achievers, but they had no particular prior knowledge of astrophysics.

### B. Student scores

#### 1. Performance data

In total, the students were able to score 42 points, 32 on the 14 HRD items and 10 on the five CFD items. A descriptive analysis shows that the students achieved an average score of $M = 22.1$ ($SD = 3.9$) on the HRD items (69% of maximum) and $M = 8.2$ ($SD = 0.8$) on the CFD items (83% of maximum). The HRD scores range from 14 (44% of maximum, 1 person) to 31 (97% of maximum, 1 person). For the CFD items, the scores range from 7 (70% of maximum, 6 persons) to 10 (100% of maximum, 2 persons). This is also shown in a histogram in Fig. 5. Overall, the students achieved a mean total score of $M = 30.5$ ($SD = 4.10$), which is 73% of the maximum.

For each item, the item difficulty (i.e., the average score obtained by all students divided by the maximum possible score) was determined, which consequently can range from 0.00 (most difficult items) to 1.00 (very easy items). The histogram in Fig. 6 shows the results in descending order of item difficulty for the HRD items, ranging from 1.00 (T5) to 0.30 (T8†). In addition, this figure includes the item difficulty of the five isomorphic items. Comparing the five HRD–CFD item pairs (T1†, T2†, T6†, T7†, and T8†) with a paired $t$ test, it can be seen that the HRD part ($M = 0.58$, $SD = 0.18$) was more difficult than the CFD part ($M = 0.83$, $SD = 0.08$), with large effect [$t(34) = -7.65$, $p = 0.000$, $d = 1.29$]. Item-wise comparisons reveal the largest difference at T6† ($d = 1.14$), followed by T8† ($d = 1.00$), T7† ($d = 0.46$),

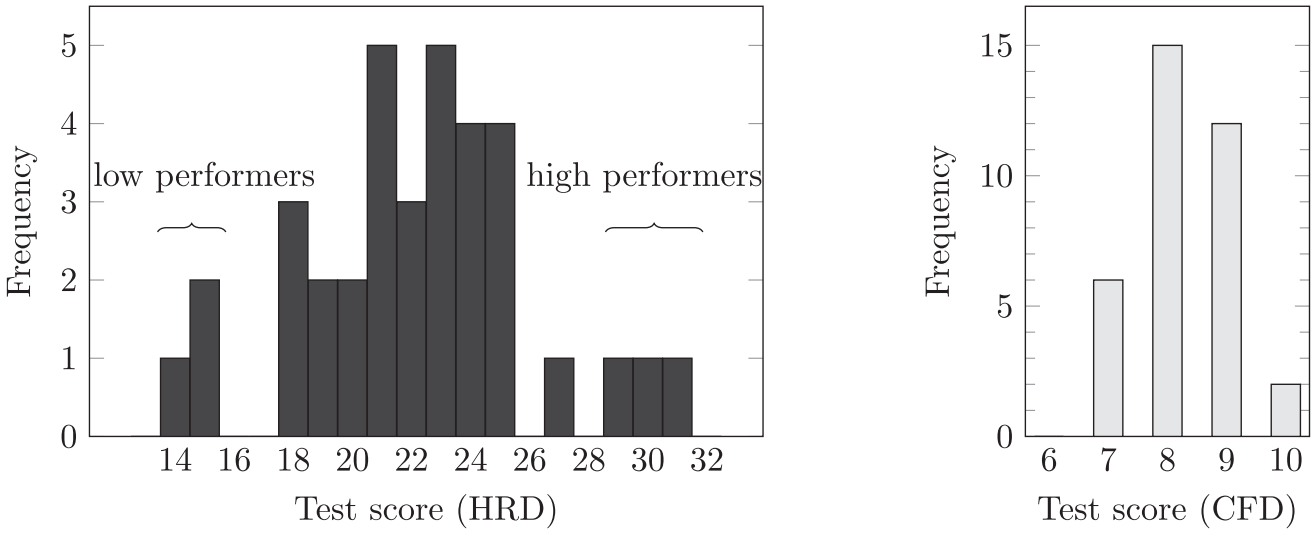


FIG. 5. Histograms of the achieved test scores for the HRD items (left, maximum score 32) and the CFD items using an $x$-$y$ diagram (right, maximum score 10). Scores lower than 13 (HRD) and 6 (CFD) were not found for the items and are therefore not shown for clarity. The three students with the lowest and highest HRD scores are designated as extreme performance groups (RQ3).

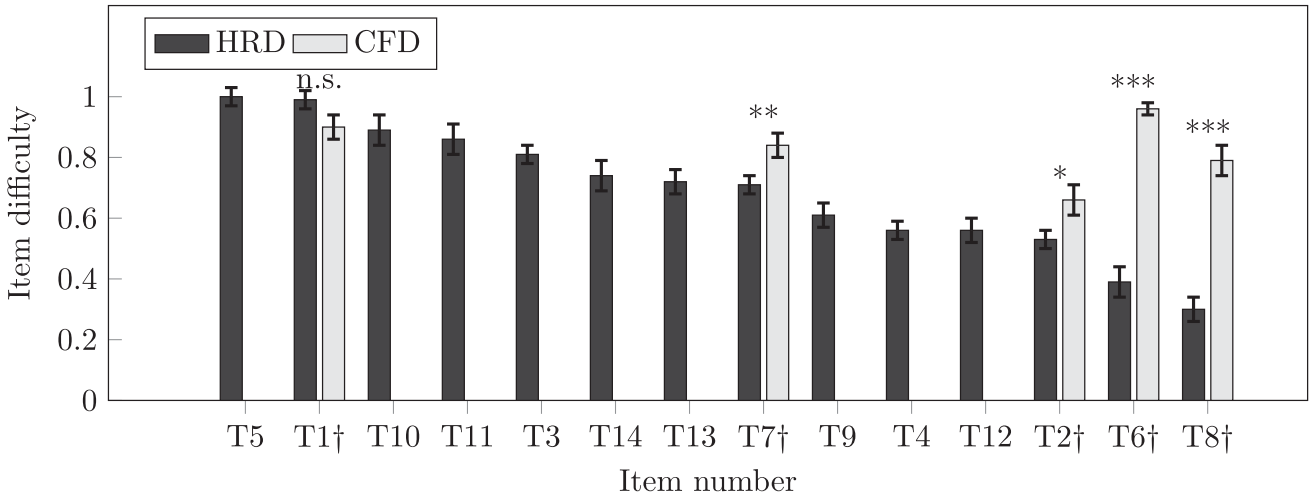


FIG. 6. Histogram of the item difficulties for all items used. The 14 HRD items (T1–T14) are ordered by item difficulty from easy (left) to difficult (right). For the items marked with the † symbol, there exists an isomorphic item with an *x*-*y* diagram. Error bars indicate the standard errors of the mean values.

and T2† ($d = 0.35$). Because of a ceiling effect, there were no differences at T1†.

#### 2. Perceived item difficulties and cognitive load

The students estimated the difficulty of each HRD item ($\alpha = 0.72$). Like item difficulty, perceived item difficulty ranges from 0.00 (difficult) to 1.00 (easy). The average perceived item difficulty is $M = 0.67$ ($SD = 0.18$). The students perceived item T2† as the easiest ($M = 0.87$, $SD = 0.19$) and item T13 as the most difficult ($M = 0.34$, $SD = 0.25$). Analysis at the item-level shows that perceived task difficulty correlates with item difficulty for only two of the HRD items (T7† and T8†). In both cases, a high value of the perceived item difficulty is correlated with a low value of the actual item difficulty—in T7† with a medium correlation ($r = -0.40$, $p < 0.05$, $N = 35$) and in T8† with a strong relation ($r = -0.60$, $p = 0.000$, $N = 35$). Thus, the easier these items were estimated to be, the poorer they were actually solved. For the other 12 items, no relationship was found between perceived item difficulty and item difficulty. This reflects an estimation bias that will be further investigated below.

In addition to the students' perceived difficulty of each item, we assessed the ICL and the GCL after completing both sets of items using three items each ($\alpha \geq 0.80$). Figure 7 shows the comparison of these diagram types (paired t-test). Regarding the HRD, both the ICL ($M = 0.51$, $SD = 0.08$) and the GCL ($M = 0.71$, $SD = 0.08$) are quite high. Concerning the CFD, the ICL ($M = 0.09$, $SD = 0.00$) and the GCL ($M = 0.29$, $SD = 0.15$) are rather low. It can be seen that the students reported a higher ICL when dealing with the HRD than when dealing with the CFD [$t(34) = 14.68$, $p = 0.000$, $d = 2.48$]. At the same time, the GCL related to the HRD is higher than that related to the CFD [$t(34) = 10.26$, $p = 0.000$, $d = 1.74$]. Both differences have a strong effect.

#### 3. Correlations

The following 10 variables were included in the correlation analysis (Pearson): grade on high school diploma, score for the HRD items, score for the CFD items, prior knowledge, SST score, mean rated perceived item difficulty, ICL (HRD), GCL (HRD), ICL (CFD), and GCL (CFD). After a Bonferroni correction with $n = 10$, no significant correlation was found at any point.

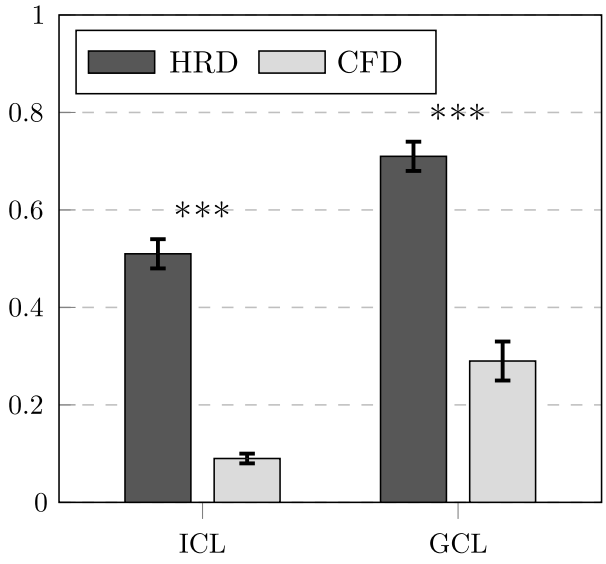


FIG. 7. Results for intrinsic cognitive load and germane cognitive load in a comparison of dealing with the HRD and an isomorphic *x*-*y* diagram. Error bars indicate the standard errors of the mean values.

| Item | AL | AR | AT | AB | i | LL | LR | LT | LB |
|---|---|---|---|---|---|---|---|---|---|
| T1 | | | | 1 | 3 | | | | 2 |
| T2 | | 3 | 2 | | | | 1 | | |
| T3 | | | 1 | 2 | | | | | 3 |
| T4 | | | 1 | 2 | | | | | 3 |
| T5 | | | 1 | | 2 | 3 | | | |
| T6 | 2 | | | | | 1 | 3 | | |
| T7 | | 1 | | | 3 | | 2 | | |
| T8 | | 1 | | | 3 | | 2 | | |
| T9 | 3 | 1 | | | | | 2 | | |
| T10 | 2 | | | | | 1 | 3 | | |
| T11 | | 2 | | | | 3 | 1 | | |
| T12 | | 2 | 3 | | | | 1 | | |
| T13 | | 3 | 2 | | | | 1 | | |
| T14 | | | 2 | | 1 | | 2 | | |

FIG. 8. Listing of the three local AOIs that receive the most attention per item. For this purpose, the TVD was normalized to the size of the AOI. Cells colored in gray indicate the AOIs that are addressed by the item text. This consists of the AOIs of the axes (A) and labels (L) in all four directions (L)eft, (R)ight, (T)op, and (B)ottom, as well as the AOI for the interior of the diagram (i).

## C. Eye-tracking data

### 1. Areas of interest and total visit duration

For the following analysis, the AOIs were used as defined in Fig. 4. For each local AOI, the total visit duration (TVD) was determined and normalized to the size of the AOI (measured by pixels) to identify the elements of the HRD that received the most attention. Figure 8 shows those three AOIs (ordered from 1 to 3) per item whose normalized TVD is the largest. In addition, table cells colored in gray indicate the AOIs belonging to the addressed physical quantities of an item. For example, in the text of item T1 the spectral class is addressed and therefore the table cells of AB-AOI and LB-AOI are colored in gray. The table cell of the interior of the diagram (i-AOI) is colored in gray only for the items T12 to T14. These items relate to stellar evolution and luminosity classes, giving much more information in the interior compared to the other items (Table IV). The AOIs are arranged so that the axes are in the first four columns, the labels are in the last four, and one column in between covers the interior. This creates symmetry with respect to the gray table cell color in Fig. 8. It is noticeable that although the LT-AOI (label of temperature) is addressed in eight items, it is never among the AOIs with the most attention. The i-AOI received high attention in four items, although it is not one of the addressed areas. The LL-AOI and the LR-AOI (labels of magnitude and luminosity) received attention in two items each, although they are not addressed. The same is true for one item for the AT-AOI (temperature axis). Other than that, the addressed and high-attention AOIs are consistent.

### 2. Comparison of HRD and CFD item pairs

In Table V, five of the HRD items are compared with the isomorphic CFD items regarding TVD and total visit counts (TVCs) using paired $t$ tests. For significant differences, the effect size $d$ was determined (Bonferroni correction by $n = 18$). Overall, the students viewed the HRD longer than the CFD (TVD of D-AOI) with large effect ($d = 0.94$). An item-wise comparison of TVD for the global AOIs between the HRD and CFD items shows that the students viewed the HRD longer than the isomorphic diagram (D-AOI), with T1† having a large effect ($d = 0.84$) and T6† having a medium one ($d = 0.65$). As for the T-AOI, an item-wise comparison does not show a consistent result for all five pairs. Here, TVD is higher for both the HRD item T6† ($d = 0.62$) and the CFD item T2† ($d = 0.68$). The item-wise comparison of the TVC for the T-AOI shows a difference at T6† ($d = 0.69$), where the HRD text was reread more frequently with medium effect. Beyond that, no differences were found.

Looking at the local AOIs, the TVD of the four axes (A) and labels (L) is compared pairwise to identify areas that were viewed for a particularly long time (Bonferroni correction by $n = 12$). Therefore, the TVD is normalized to the time spent on the item. For the T4† and T5† pairs, the luminosity axis and label were viewed longer than the $y$ axis and its label, with medium to strong effects [$t(34) \geq 3.27$, $p \leq 0.024$, $d \geq 0.55$]. This is also evident from Fig. 9, which contrasts the heat maps of T8†, providing a qualitative comparison. Heat maps provide

TABLE V. Comparison of the HRD and CFD item pairs in terms of total visit duration (TVD) and total visit counts (TVCs) of the D- (diagram) and T-AOIs (text). The mean value $M$ and the standard deviation in parentheses ($SD$) are given. For significant differences, the effect size $d$ is given.

| | Total | | | T1 | | | T2 | | | T6 | | | T7 | | | T8 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HRD | CFD | $d$ | HRD | CFD | $d$ | HRD | CFD | $d$ | HRD | CFD | $d$ | HRD | CFD | $d$ | HRD | CFD | $d$ |
| TVD of D-AOI | 76.35 (40.99) | 51.50 (25.97) | 0.94 | 16.70 (8.03) | 10.84 (6.70) | 0.84 | 11.34 (5.87) | 8.25 (6.50) | ⋯ | 13.33 (10.91) | 6.90 (5.84) | 0.65 | 13.57 (11.50) | 9.43 (5.49) | ⋯ | 21.41 (18.80) | 16.08 (11.94) | ⋯ |
| TVD of T-AOI | 25.11 (16.90) | 18.51 (10.68) | ⋯ | 1.86 (1.36) | 1.17 (1.24) | ⋯ | 3.31 (2.44) | 5.95 (3.92) | 0.68 | 12.99 (13.26) | 4.64 (4.59) | 0.62 | 3.20 (2.98) | 2.13 (1.49) | ⋯ | 3.74 (2.77) | 4.62 (2.99) | ⋯ |
| TVC of T-AOI | 17 (8) | 16 (7) | ⋯ | 2.86 (1.44) | 2.00 (1.63) | ⋯ | 3.40 (1.83) | 4.31 (2.92) | ⋯ | 5.54 (4,64) | 2.40 (1.82) | 0.69 | 2.37 (1.70) | 3.29 (1.53) | ⋯ | 2.77 (1.68) | 3.74 (2.06) | ⋯ |

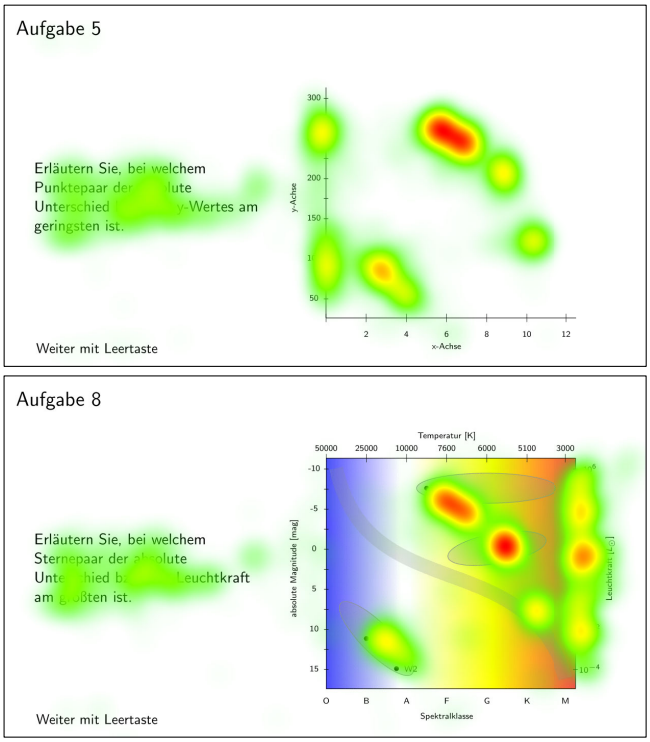


FIG. 9. Heat maps of an isomorphic pair of items (T8†) as originally used in the study (English translations can be found in the Supplemental Material [51]). These heat maps are for all participants. Information must be taken from the left $y$ axis (top) and from the right luminosity axis (bottom). In both cases, the relevant values of the points or stars must be extracted and compared. Note the more extensive consideration of the logarithmic luminosity axis (bottom) compared to the linear $y$ axis (bottom).

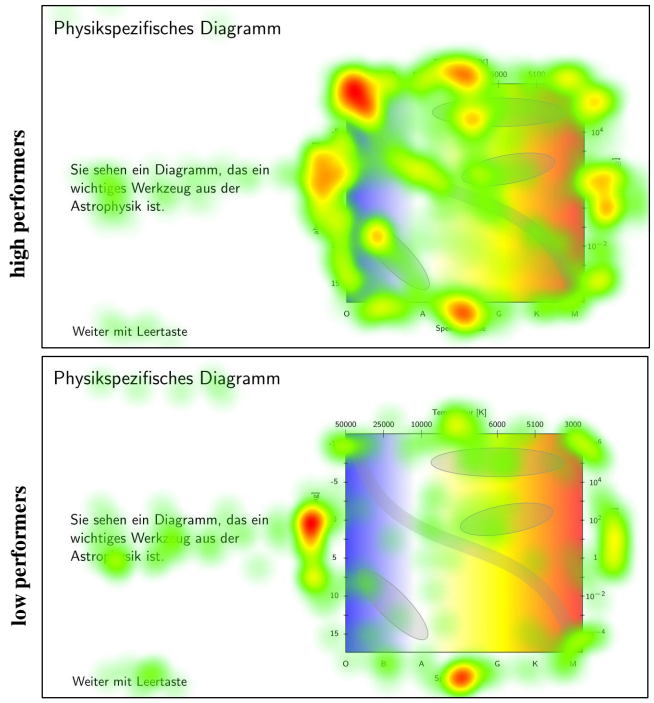


FIG. 10. Heat maps of all high performers (top) and all low performers (bottom) in comparison. The heat maps refer to the stimuli as originally used in the study (English translations can be found in the Supplemental Material [51]) where students were learning about the HRD for the first time (without problem solving).

an overview of the distribution of visual attention on a stimulus. The most intensively observed areas are highlighted in red. The visualization can simultaneously display data from several viewers. Figure 9 indicates that the students fixated mostly on the entire luminosity axis and not just on smaller, selected areas, such as within the $y$ axis. They also focused more on the luminosity label than on the $y$ label. Consequently, the quantitative (ET metrics) and qualitative data (heat maps) come to the same result for this pair. Similar agreement is also seen for the heat maps of the other pairs, but they are not shown here. Moreover, the quantitative results show that the magnitude label in T3† [$t(34) = 4.12, p = 0.000, d = 0.70$] and the spectral class label in T1† [$t(34) = 3.09, p = 0.048, d = 0.52$] were considered longer than the $y$ or $x$ label, with medium effects. No quantitative differences were found between the item pairs with respect to the other axes and labels.

### 3. High versus low test score

Two subgroups were formed by a *post-hoc* split of the test score, (i) *high performers* with a score ≥29 (>90% of maximum) and (ii) *low performers* with a score ≤15 (<50% of maximum). Both groups consisted of three students whose scores differed significantly from the scores of the other students (Fig. 5). Note that the student with a score of 27 did not belong to the high-performers group; the group sizes are therefore the same. On average, the total time taken to solve all HRD items (phase 2a) was $M = 1615$ s ($SD = 249$ s) for high performers and $M = 833$ s ($SD = 197$ s) for low performers. Thus, high performers took longer, with a strong effect ($d = 3.48$).

A qualitative comparison of visual attention using heat maps was conducted to determine the difference between these two student groups when dealing with the HRD. Two examples are presented, each posing different requirements to the students. Figure 10 shows heat maps for the students' first encounter with the HRD, in which they were to learn about the diagram for the first time. Therefore, there was no task setting or problem solving here. In contrast, Fig. 11 shows heat maps for item T6 as an example of problem solving, where the brightest of three given stars must be identified by its magnitude (Table IV). Building on both figures, the qualitative results are as follows:

- In general, low performers focused on fewer elements of the HRD.
- Low performers considered the four axes and the four luminosity classes much less comprehensively than high-level students.

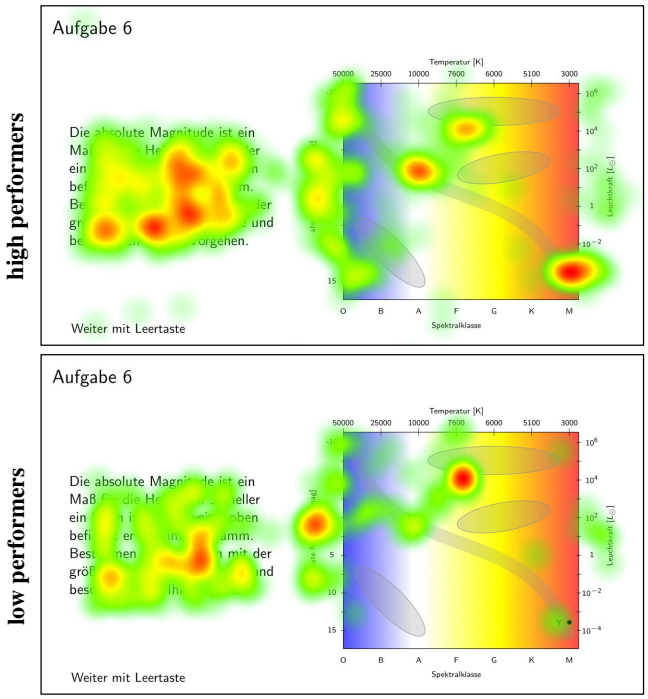


FIG. 11. Heat maps of all high performers (top) and all low performers (bottom) in comparison. The heat maps show the visual attention when solving item T6 as originally used in the study (English translations can be found in the Supplemental Material [51]). Here, three stars are given, and the problem requires comparing their magnitude values to identify the brightest one (bottom right).

- During problem solving, low performers fixated less along the entire magnitude axis and considered only its upper region. However, the high performers fixated on the entire axis and were thus able to identify the star they were looking for (highest magnitude) as the lowest in the diagram.

A qualitative comparison of the other items shows similar results. The heat maps of the other items can be found in the Supplemental Material [51]. The results suggest that high performers looked at the elements of the HRD more carefully and in more detail. In particular, they looked at the axes as a whole and did not just fixate on individual, isolated areas of the axes. Quantitative differences, for example, in terms of TVD and TVC, were not found.

#### *4. Correlation analysis including visual attention*

A correlation analysis shows the relationships between student ratings, scores, and eye-tracking data. First, the higher the perceived item difficulty of the HRD items was evaluated by the students, the more time students spent on the diagram (D-AOI) [$r = 0.72, t(13) = 5.99, p = 0.000$] and on the HRD text (T-AOI) [$r = 0.87, t(13) = 10.22$, $p = 0.000$]. Consequently, the students associated processing time with item difficulty. Second, an analysis of TVCs for the T-AOI shows that the students reread the HRD text more frequently as perceived item difficulty increased [$r = 0.91$, $t(13) = 12.76, p = 0.000$]. Further analysis examined the correlation between the HRD score and the TVD for the nine local AOIs. A longer total recording is associated with a higher score [$r = 0.70, t(34) = 5.54, p = 0.000$]. Furthermore, the higher the students' score, the higher the absolute TVD on the axes (A-AOIs) [$r = 0.61, t(34) = 4.35, p = 0.000$] and on the interior part of the diagram (i-AOI) [$r = 0.64$, $t(34) = 4.71, p = 0.000$]. This relationship also holds for the TVD normalized to the total recording. In summary, the students with higher scores paid more attention to the axes and the interior part of the diagram than the students with lower scores. As for the text (T-AOI), the score does not correlate with the TVD.

### D. Student interviews

Table VI summarizes the learning difficulties in dealing with the HRD based on a qualitative analysis of the interviews. A total of 233 statements about learning difficulties are included in the analysis, so that on average a student reported $M = 6.66$ ($SD = 2.4$) problems. More than half of these 233 learning difficulties relate to the four physical quantities (luminosity, absolute magnitude, temperature, and spectral class) represented in the HRD. However, in the subsequent quantitative analysis of the interview data, each category is counted only once per person, even if it was coded multiple times throughout the transcript. The only interest here is whether a person commented on a learning difficulty or not. It is not of interest how many times they expressed it (see Table VI).

The main category *physical meaning* occurred most frequently. Here, the students reported learning difficulties due to not knowing or not understanding the term, definition, or interpretation of a physical quantity and its properties (e.g., unit). This difficulty occurred with all four dimensions shown in the HRD (specification). Absolute magnitude was mentioned by most students (77.1%), followed by spectral class (45.7%) and luminosity (31.4%), while temperature was mentioned here by only one student. The two main categories *axis scaling* and *axis orientation* also account for common learning difficulties. The axis orientation of the absolute magnitude was found to be difficult by 65.7% of the students, and about 17.1% reported difficulties with the nonlinear scaling. Regarding the temperature axis, the orientation (42.9%) and the nonlinear scaling (37.1%) were perceived as challenging. In addition, the logarithmic scaling of the luminosity axis (37.1%) and the interval scale of the spectral class (25.7%) were identified as difficulties.

Many students (62.9%) referenced the differences between the HRD and the familiar Cartesian coordinate system, such as two axes and the origin as a point on the coordinate cross. *Complexity* refers to learning difficulties caused by the abundance of information and general complexity of the diagram and was mentioned by 54.3%

TABLE VI. Summary of perceived learning difficulties in using the HRD. A category was counted only once per student, even if multiple statements were made about it. Thus, the count of students (out of 35) who commented on a category is provided.

| Main categories (specification) | Students | Statement example |
|---|---|---|
| Physical meaning ($L, M, T, SC$) | 30 | What magnitude is, I can't imagine at all. |
| Axis orientation ($M$, $T$) | 26 | What I stumbled on first [...] is this thing about temperature being the other way around than you would expect. |
| Axis scaling ($L, M, T, SC$) | 22 | Surprisingly, I felt that the luminosity was not given linearly, but grew logarithmic, [...] it was not as usual. And the same applies to the temperature, that these two have not always increased the same value, but logarithmic. |
| Diagram characteristics (origin, 4 axes) | 22 | [...] that one cannot identify a distinct origin. |
| Complexity | 19 | [...] because I got much more information than I needed in the end, and then I also thought about it. |
| Physical relationships | 15 | But getting the relationships between the axes, that was more difficult. |
| Luminosity classes | 15 | I was a bit confused by the fact that one area is not indicated as an oval but as a curve [comment: student refers to the main series], and that caused a bit of confusion. |
| Color (background, $T$-relation) | 14 | The colors initially confused me because of the temperature. Towards red is rather hot, towards blue-purple it's rather cold. And then I realized only with time that this is total nonsense and that the temperature grows the other way around. |
| Invisible aspects (radius, mass) | 4 | [...] and the radius is also not shown here at all. That's also the point. You just had to think about that in addition. [...] That was always a difficulty. |
| Stellar evolution | 4 | I must say that I felt that the description of [evolution] tracks in this diagram, especially from the Sun, was quite difficult, because you had curves that were just not straight, but jumped around [discontinuous]. |

of the students. Furthermore, 42.9% of the students stated they had difficulties regarding the conceptual understanding of *physical relationships* presented in the HRD (e.g., $L \propto TR$). Regarding the visual representation and meaning of the *luminosity classes*, 42.9% of the students stated they had difficulties. Learning difficulties related to the *color* of the background and its relationship to temperature were experienced by 40.0% of the students. The main category of *invisible aspects*, which deals with the missing plot of stellar radius and mass, plays an even smaller role, affecting only 11.4% of the students. In addition, 11.4% of the students had difficulty with stellar evolution as the physical meaning of the HRD and the corresponding representation of time as another dimension in the diagram.

The students commented not only on the learning difficulties mentioned but also on the strategies they used when working with the HRD. These strategies are divided into those that were used during problem solving and those that were used one step ahead when becoming familiar with the diagram. When becoming familiar with the HRD, some students (12) first made sense of the relationships between the two parallel axes (luminosity and magnitude, temperature and spectral class) and then used this comprehension for problem solving. Furthermore, 13 students stated that looking carefully at the axes is important and helpful in understanding the diagram. Next, more strategies came into use during the problem-solving process. The most frequently mentioned strategy (20) is that of identifying task-relevant information and hiding task-irrelevant information. In addition, some students explicitly ignored the color (15) or the representation of luminosity classes (4). Further information reduction occurred for seven students by considering only one horizontal and one vertical axis (e.g., luminosity and temperature) instead of all four axes. Some of the students (5) even reported ignoring the axes and orienting themselves spatially in the diagram when solving problems (e.g., the more to the top, the greater the luminosity). A total of six students indicated that they either imagined a Cartesian coordinate system and connected it directly to the HRD or copied their procedure from phase 1a in dealing with the $x$-$y$ diagram directly to phase 2a (HRD). The last strategy in particular may explain some of the problem-solving mistakes and learning difficulties, which will be covered in more detail in the discussion. Although the additional color in the background of the HRD can cause difficulties and one strategy mentioned is to ignore it, it can also serve as a kind of guideline in the vertical direction. For example, some students (7) compared the color scheme to a coordinate grid, where the color gradient provides orientation in the vertical direction.

## VI. DISCUSSION

In this study, we used eye tracking, retrospective interviews, and various questionnaires to investigate how learners extract information from the HRD. The research interest was to identify learning difficulties in using the HRD. Thirty-five physics students participated voluntarily.

They had no specific prior astrophysics knowledge but they can be characterized as high achievers. In the following, the results will be discussed according to the three research questions.

### A. Extracting information from the HRD

The first research question reads *How well do students succeed in extracting information from the HRD and what difficulties do students report?* The HRD items were designed to be solved correctly without specific prior knowledge, and the test scores students achieved confirms this. The students achieved between 44% and 97% of the maximum score, and the full score was credited at least once for each item. The 14 HRD items had increased requirements in terms of using a diagram (Table IV). As Fig. 6 shows, this increase in item requirements cannot be supported by item difficulty (determined by students' scores). Items T6† and T8†, although among the first items with lower requirements, were poorly solved by the students (item difficulty $< 0.5$). In contrast, T10 and T11, both with medium requirements, were solved quite well (item difficulty $> 0.8$).

The fact that item difficulty was not as expected (T1 as easy to T14 as difficult) requires a closer look at the individual HRD items. First, the two well-solved items are considered. Both items T10 and T11 expect a qualitative statement about relationships between two plotted quantities. Probably this was easier for the students than expected, as no quantitative analysis of the individual scales was necessary. Second, the two poorly solved items are considered. In item T6†, the magnitude scale must be used to determine the brightest star, that is, the star with the lowest value. Here, the students with an incorrect solution identified the star with the largest magnitude as the brightest one. In item T8†, the pair of stars with the largest absolute difference in luminosity had to be identified. Note, this is not the pair with the largest spatial distance in the diagram (which reflects the typical incorrect answer) due to the logarithmic scale. For both HRD items, there were isomorphic CFD items for each of which the students performed better compared to the HRD items, with large effect. As the perceptual actions of information extracting are the same for each pair, we suggest that the HRD itself is the source of these differences. As said in the literature, there may be a difficulty with the historical definition of magnitude and the logarithmic scale of luminosity [1]. This assumption is supported by our results. It is also interesting that the easier T8† was evaluated by the students to be, the poorer it was actually solved. We assume that rating T8† as easy and solving it incorrectly at the same time is due to not recognizing the logarithmic scale and the difficulty this implies for comparisons. Therefore, the students underestimated the difficulty in the mistaken belief that they have solved the item correctly.

Overall, the comparison of the item difficulties of all five isomorphic pairs shows that the CFD items were solved more successfully with large effect. For T2†, it can be assumed that this is due to the unfamiliar orientation of the temperature axis [1]. The difference at T7† also supports the suggestion that the logarithmic luminosity axis is a learning difficulty.

The scores already provide good evidence of learning difficulties, which is supported by the additional data obtained from the interviews. Based on this, the magnitude axis orientation, the nonlinear luminosity scale, and the temperature axis orientation were challenging for more than one-third of the students. Thus, these results support the assumptions based on the performance data. Furthermore, 10 main categories were found that can be used to name the mentioned learning difficulties. As a link back to the literature, the learning difficulties uncovered here are mapped into the four potential barriers according to Airey and Eriksson [1].

- *History:* The main category *physical meaning* can be placed here in terms of magnitude and spectral class. (The stellar quantities luminosity and temperature have no historical roots, so these specifications are not classified here.)
- *Omission:* This category matches the main category *invisible aspects*.
- *Overloading:* This category matches with the main category *complexity* and the specification *background color*. In addition, the main category *luminosity classes* is also suitable here.
- *Expectations:* The main categories *axis orientation* and *axis scaling* can be located here. Note that these also include magnitude and spectral class, which for [1] is included in *History*. Furthermore, the color specification *T relation* matches here. This barrier can also be properly extended by our category *diagram characteristics*, as this refers to the unfulfilled expectations of the representation form. (The *life-death analogy* does not exist in our categories.)

The main categories *physical relationships* and *stellar evolution* and the specifications *physical meaning of luminosity* and *temperature* cannot be mapped. They have in common that they refer to the astrophysical content. Learning difficulties associated with these categories are related to understanding, interpreting, and identifying relationships in stellar physics. Based on this, one more potential barrier can be added to the list of four—the *content*. In our study, *expectations* and *overloading* are the barriers that most students struggled with. *History* and *content* also caused serious learning difficulties, while *omission* hardly seemed to be challenging at all.

Finally, the measurement of CL when dealing with the HRD and the isomorphic CFD is discussed. These data also support the finding that dealing with the HRD is challenging for learners. Compared to the CFD, the ICL and the

GCL are higher for HRD items with large effects. Accordingly, the students perceived the learning content and information in the HRD (refers to ICL [36,38]) as much more complex, which was also shown in the interviews. The ICL we measured is higher for the HRD tasks, which means that the intrinsic task demands on these items were higher for our sample. This may be due to the high element interactivity that determines ICL [36,38,39], as the understanding of *physical relationships* was identified as a learning difficulty in the interviews. The interview analysis shows that the students used some strategies to reduce the intrinsic complexity of the HRD and the CL. They used approaches consistent with the *information-reduction hypothesis* [45,46] and *chunking* [40]. In addition, building on the scale for GCL [52], it can be concluded that the students perceived a higher learning effect and better understanding when working on the HRD items compared to the CFD items. We attribute the fact that the GCL is higher in the HRD tasks to the inferior cognitive challenge in the CFD tasks; the procedures that were required there without a physical context were already known to our sample and so our sample could not learn anything new in the process. It should also be noted that our study was concerned with investigating problem solving, which means that learning was not the primary intention at all. Nevertheless, the progressive structure of the tasks allowed students to learn as they progressed through the test, which is reflected in an increased expression of GCL. It is an interesting finding that the reported total load across tasks is not a constant, i.e., the sum of all partial loads is not always constant: one can have both less ICL and less GCL on one type of task than on another, and that is the case here because the demands are so different.

### B. Visual attention while using the HRD

The second research question reads *Can the reported difficulties while working with the HRD be empirically supported by the analysis of visual attention?* The results in Fig. 8 suggest that the students' visual attention was generally distributed in accordance with item requirements. Furthermore, the analysis of visual attention empirically supports the finding that the HRD items were more difficult than the CFD items. First of all, this manifests in the fact that, overall, the HRD was viewed for a longer time than the isomorphic diagram, indicating higher CL [55]. In addition, the following HRD areas received more attention than their isomorphic counterparts: luminosity axis, luminosity label, magnitude label, and spectral class label. Referring to other eye-tracking studies [8,10], it is reasonable to assume that these axis labels received more attention because they were unfamiliar to the students. This also explains why no special attention was paid to the temperature label, which can be seen as a familiar physical quantity for physics students. The analysis of TVD normalized to the size of each local AOI also showed that the temperature label did not receive special attention for any of the 14 HRD items (Fig. 8). In contrast, the other three labels received special attention, in some cases even when the physical quantity was not addressed at all.

For all 14 HRD items, the axes addressed coincide with the AOIs that received the most attention in general (Fig. 8). Therefore, no conclusion about difficulties with individual axes can be made based on this quantitative data. However, the fact that the luminosity axis received more attention than the isomorphic $y$ axis is also evident in the qualitative heat map data (Fig. 9). Based on the results regarding the first research question, we can assume that this is due to the difficulty of this axis. However, based on the eye-tracking data, such statements cannot be made regarding the other axes.

### C. Comparing high vs low performers

The third research question reads *How does the level of expertise impact processing of the HRD?* The correlation analysis shows that the more successful students are in problem solving, the more attention they pay to the axes and the interior of the HRD. Furthermore, a high score correlates with a longer total time to solve all items. In addition to this analysis of all participants, two special subgroups were compared. Using a *post hoc* split of the test score, we identified two groups as high and low performers (best three vs bottom three students). The heat maps show significant differences in their visual attention that—together with the other data sources—allow the formulation of the following heuristics, serving as successful problem-solving approaches to the HRD:

- *Become familiar*: High performers take time to become familiar with and understand the diagram and its plotted quantities even before problem solving.
- *Accuracy*: High performers look at the individual axes in detail, paying particular attention to orientation and scaling, analyzing the entire axis (before and during problem solving).
- *Adaption*: High performers become aware of the unfamiliar properties of the diagram and enable rethinking of the heuristics that were useful for previous diagrams (before and during problem solving).
- *Relationships*: High performers become aware of the relationships between the quantities presented (before and during solving the problem).

It may happen that some students bluntly transferred the procedure from the $x$-$y$ diagrams they first worked on to the HRD. This is obviously wrong and indicates that only superficial commonalities of the tasks were considered without understanding their underlying structure. In this context, too much focus on spatial orientation (left, right, top, or bottom) instead of a systematic inspection of the axes can also lead to problems when procedures are mechanistically copied from typical $x$-$y$ diagrams. However, it requires cognitive effort and training to

overcome familiar strategies, such as the culturally conditioned selection of information from Cartesian coordinate systems, as they are deeply rooted in our routines [32].

This becomes particularly evident in item T6†, which has been discussed above (cf. Fig. 11 for heat maps). Here, all low performers incorrectly identified the top star as the one with the largest magnitude, arguing about the spatial position in the diagram. Consequently, the step of rethinking familiar strategies is very important for the successful extraction of information from the HRD.

## VII. LIMITATIONS AND DESIDERATA

A limitation of this study is the CFD format, that is, the *x*-*y* diagram used. We started benchmarking the HRD against the simplest form of a diagram (consisting of two linear axes) for a good reason, which is that what most students know. In a further study, the context-free diagram could be modified systematically to mimic the layout of the HRD in more aspects. To this end, four axes or nonlinear axis scaling could be used to increase the similarity of the isomorphic pairs.

Given the learners' high level of prior achievement in school, it is an open question whether the assumption holds that they apply everyday heuristics to the color scale, as Airey and Eriksson hypothesized [1]. After all, in some chemical and physical contexts (e.g., gas flame and light-emitting diode) red color is related to low temperature and blue color to high temperature (color temperature scale)—the opposite of the everyday heuristic that red is associated with a high temperature. A further study could survey the students' familiarity with the scientific understanding of color to investigate its influence on interpreting the HRDs color scheme.

With 35 participants, the sample size of this eye-tracking study is larger than the average sample size in mathematics education research ($M = 28.56$, $SD = 21.70$ participants) [56] but smaller than the average in PER ($M = 54.4$, $SD = 29.9$ participants) [12]. Larger groups are desirable to improve the statistical robustness of group comparisons, particularly when comparing high- and low-performing students.

## VIII. CONCLUSION

In summary, this study identified learning difficulties when novice learners used the HRD, provided insight into how students engage with the HRD when completing tasks, and revealed what strategies students use. Most of the results are consistent with the literature, corroborating and extending the research about the pedagogical affordances of the HRD. Furthermore, we were able to reveal various difficulties using different measurement tools (scores, interviews, and gaze data) and how persistent they are.

Based on student interviews and data on attention allocation, it was possible to derive initial statements that serve as heuristics for successful problem solving with the HRD. These consist of four approaches that can be used as research-informed takeaways in teaching-learning situations to introduce the HRD—become familiar, be accurate, adapt, and study relationships.

Using product-oriented data sources (students' scores and ratings), procedural data sources (visual attention), and think-aloud data enabled us to study the students' use of the HRD on multiple levels, uncovering the learning processes beyond the learning difficulties. As a by-product of this work, it is worth pointing out that the students actually experienced learning gains when completing the experimental procedure, even though learning was completely self-paced with no guidance by a teacher or instructor. Therefore, we feel encouraged to incorporate the study material (i.e., the items) into future material development, for example, supporting astrophysics tutorials.

In a next step, the study results can serve the development of (interactive) learning materials for the HRD, introducing learners to the characteristics of the HRD step by step and addressing the difficulties explicitly. Because of its high disciplinary affordance and its low pedagogical affordance, work like this is required to increase pedagogical affordance and effectively support learners in using the diagram.

[1] J. Airey and U. Eriksson, Unpacking the Hertzsprung-Russell diagram: A social semiotic analysis of the disciplinary and pedagogical affordances of a central resource in astronomy, Designs for Learning **11**, 99 (2019).

[2] E. Brogt, Pedagogical and curricular thinking of professional astronomers teaching the Hertzsprung-Russell diagram in introductory astronomy courses for non-science majors, Ph.D. thesis, The University of Arizona (2009).

[3] U. Eriksson, M. Rosberg, and A. Redfors, Disciplinary discernment from Hertzsprung-Russell-diagrams, *Nordic Research Symposium on Science Education, Trondheim* (2017), https://nfsun.org.

[4] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

[5] G. Leinhardt, O. Zaslavsky, and M. K. Stein, Functions, graphs, and graphing: Tasks, learning, and teaching, Rev. Educ. Res. **60**, 1 (1990).

[6] M. C. Linn, J. W. Layman, and R. Nachmias, Cognitive consequences of microcomputer-based laboratories: Graphing skills development, Contemp. Educ. Psychol. **12**, 244 (1987).

[7] T. Wemyss and P. Van Kampen, Categorization of first-year university students' interpretations of numerical linear distance-time graphs, Phys. Rev. ST Phys. Educ. Res. **9**, 010107 (2013).

[8] A. Susac, A. Bubic, E. Kazotti, M. Planinic, and M. Palmovic, Student understanding of graph slope and area under a graph: A comparison of physics and non-physics students, Phys. Rev. Phys. Educ. Res. **14**, 020109 (2018).

[9] A. Susac, A. Bubic, M. Planinic, M. Movre, and M. Palmovic, Role of diagrams in problem solving: An evaluation of eye-tracking parameters as a measure of visual attention, Phys. Rev. Phys. Educ. Res. **15**, 013101 (2019).

[10] P. Klein, S. Küchemann, S. Brückner, O. Zlatkin-Troitschanskaia, and J. Kuhn, Student understanding of graph slope and area under a curve: A replication study comparing first-year physics and economics students, Phys. Rev. Phys. Educ. Res. **15**, 020116 (2019).

[11] C. Hoyer and R. Girwidz, Animation and interactivity in computer-based physics experiments to support the documentation of measured vector quantities in diagrams: An eye tracking study, Phys. Rev. Phys. Educ. Res. **16**, 020124 (2020).

[12] L. Hahn and P. Klein, Eye tracking in physics education research: A systematic literature review, Phys. Rev. Phys. Educ. Res. **18**, 013102 (2022).

[13] H. Draper, On photographing the spectra of the stars and planets, Am. J. Sci. **s3-18**, 419 (1879).

[14] E. Pickering, The Henry Draper Memorial 1, Nature (London) **36**, 31 (1887).

[15] K. R. Lang, *Essential Astrophysics* (Springer, Berlin, 2013).

[16] A. J. Cannon and E. C. Pickering, The Henry Draper Catalogue 0h, 1h, 2h, and 3h, Ann. Harvard College Obs. **91**, 1 (1918), https://ui.adsabs.harvard.edu/abs/1918AnHar..91....1C/abstract.

[17] E. Hertzsprung, *Publikationen des Astrophysikalischen Obsservatoriums zu Potsdam, 63* (Akad.-Verlag, 1911).

[18] H. N. Russell, Relations between the spectra and other characteristics of the stars. II. Brightness and spectral class, Nature, **93**, 252 (1914).

[19] J. C. Kapteyn, On the luminosity of the fixed stars, Publ. Kapteyn Astron. Lab. Groningen **11**, 1 (1902).

[20] R. Miles, A light history of photometry: From Hipparchus to the Hubble Space Telescope, Journal of the British Astronomical Association **117**, 172 (2007).

[21] J. Airey, Social semiotics in higher education: Examples from teaching and learning in undergraduate physics, in *Proceedings of the Concorde Hotel/National Institute of Education, Singapore* (Swedish Foundation for International Cooperation in Research in Higher Education (STINT), 2015).

[22] A. A. diSessa, What do " just plain folk" know about physics, *The Handbook of Education and Human Development: New Models of Learning, Teaching, and Schooling* (Wiley-Blackwell, 1996), pp. 709–730, 10.1111/b.9780631211860.1998.00031.x.

[23] B. W. Carroll and D. A. Ostlie, *An Introduction to Modern Astrophysics* (Cambridge University Press, Cambridge, England, 2017).

[24] G. M. Bowen and W.-M. Roth, Lecturing graphing: What features of lectures contribute to student difficulties in learning to interpret graph?, Res. Sci. Educ. **28**, 77 (1998).

[25] F. R. Curcio, Comprehension of mathematical relationships expressed in graphs, J. Res. Math. Educ. **18**, 382 (1987).

[26] S. Pinker, A theory of graph comprehension, in *Artificial Intelligence and the Future of Testing*, edited by R. O. Freedle (Routledge, London, 1990), pp. 73–126.

[27] P. Shah and J. Hoeffner, Review of graph comprehension research: Implications for instruction, Educ. Psychol. Rev. **14**, 47 (2002).

[28] S. Wineburg, J. Breakstone, S. McGrew, and T. Ortega, Why google can't save us: The challenges of our post-Gutenberg moment, in *Positive Learning in the Age of Information (PLATO): A Blessing or a Curse*?, edited by O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel (Springer VS, Wiesbaden, 2018), p. 221.

[29] B. Cowie and B. Cooper, Exploring the challenge of developing student teacher data literacy, Assess. Educ. Principles, Policy Pract. **24**, 147 (2017).

[30] S. Brand-Gruwel, I. Wopereis, and A. Walraven, A descriptive model of information problem solving while using internet, Comput. Educ. **53**, 1207 (2009).

[31] L. C. McDermott, M. L. Rosenquist, and E. H. van Zee, Student difficulties in connecting graphs and physics: Examples from kinematics, Am. J. Phys. **55**, 503 (1987).

[32] D. Chumachemko, A. Shvarts, and A. Budanov, The Development of the Visual Perception of the Cartesian Coordinate System: An Eye Tracking Study, in *Proceedings of the Joint Meeting 2–313 of PME 38 and PME-NA*, edited by C. Nicol, P. Liljedahl, S. Oesterle, and D. Allan (PME Publishing, Vancouver, CA, 2014), Vol. 36, pp. 313–320.

[33] P. Klein, J. Viiri, S. Mozaffari, A. Dengel, and J. Kuhn, Instruction-based clinical eye-tracking study on the visual interpretation of divergence: How do students look at vector field plots?, Phys. Rev. Phys. Educ. Res. **14**, 010116 (2018).

[34] P. Klein, J. Viiri, and J. Kuhn, Visual cues improve students' understanding of divergence and curl: Evidence from eye movements during reading and problem solving, Phys. Rev. Phys. Educ. Res. **15**, 010126 (2019).

[35] P. Chandler and J. Sweller, Cognitive load theory and the format of instruction, Cognit. Instr. **8**, 293 (1991).

[36] J. Sweller, Element interactivity and intrinsic, extraneous, and germane cognitive load, Educ. Psychol. Rev. **22**, 123 (2010).

[37] J. Sweller, J. J. G. Merriënboer, and F. Paas, Cognitive architecture and instructional design: 20 years later, Educ. Psychol. Rev. **31**, 261 (2019).

[38] J. Sweller, Cognitive load theory, learning difficulty, and instructional design, Learning Instruct. **4**, 295 (1994).

[39] J. Sweller and P. Chandler, Why some material is difficult to learn, Cognit. Instr. **12**, 185 (1994).

[40] F. Gobet, P. C. R. Lane, S. Croker, P. Cheng, G. Jones, I. Oliver, and J. Pine, Chunking mechanisms in human learning, Trends Cognit. Sci. **5**, 236 (2001).
[41] W. G. Chase and H. A. Simon, Perception in chess, Cogn. Psychol. **4**, 55 (1973).
[42] P. C. R. Lane, P. C. H. Cheng, and F. Gobet, CHREST+: A simulation of how humans learn to solve problems using diagrams, Artif. Intell. Simul. Behav. Quart. **103**, 24 (2000), http://bura.brunel.ac.uk/handle/2438/1356.
[43] G. A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychol. Rev. **63**, 81 (1956).
[44] J. J. G. van Merriënboer and J. Sweller, Cognitive load theory and complex learning: Recent developments and future directions, Educ. Psychol. Rev. **17**, 147 (2005).
[45] H. Haider and P. A. Frensch, The role of information reduction in skill acquisition, Cogn. Psychol. **30**, 304 (1996).
[46] H. Haider and P. A. Frensch, Eye movement during skill acqusition: More evidence for the information-reduction hypothesis, J. Exper. Psychol. Learn. Memory Cogn. **25**, 172 (1999).
[47] F. Paas and J. J. G. van Merrienboer, Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks, Curr. Dir. Psychol. Sci. **29**, 394 (2020).
[48] T. de Jong, Cognitive load theory, educational research, and instructional design: Some food for thought, Instr. Sci. **38**, 105 (2010).
[49] Kultusministerkonferenz, Abiturnoten im Ländervergleich. https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik/abiturnoten.html, Berlin (2021).
[50] P. Klein, L. Hahn, and J. Kuhn, Einfluss visueller Hilfen und räumlicher Fähigkeiten auf die graphische Interpretation von Vektorfeldern: Eine Eye-Tracking-Untersuchung, ZfDN **27**, 181 (2021).
[51] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.020121 for a full evaluation of the study quality.
[52] J. Leppink, F. Paas, C. P. M. Van der Vleuten *et al.*, Development of an instrument for measuring different types of cognitive load, Behav. Res. Methods, **45**, 1058 (2013).
[53] P. Shah and A. Miyake, The separability of working memory resources for spatial thinking and language processing: An individual differences approach, J. Exper. Psychol. **125**, 4 (1996).
[54] P. Mayring, Qualitative content analysis: Theoretical foundation, basic procedures and software solution, Klagenfurt (2014), https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173.
[55] A. Susac, A. Bubic, P. Martinjak, M. Planinic, and M. Palmovic, Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study, Phys. Rev. Phys. Educ. Res. **13**, 020125 (2017).
[56] A. R. Strohmaier, K. J. MacKay, A. Obersteiner, and K. M. Reiss, Eye-tracking methodology in mathematics education research: A systematic literature review, Educ. Stud. Math. **104**, 147 (2020).