(How) Do Pre-service Teachers Use YouTube Features in the Selection of Instructional Videos for Physics Teaching?

Philipp Bitzenbauer[1] · Tom Teußner[2] · Joaquin M. Veith[3] ·
Christoph Kulgemeyer[4]

Abstract

This mixed-methods study examines how pre-service teachers select instructional videos on YouTube for physics teaching. The study focuses on the role of surface features that YouTube provides (e.g., likes, views, thumbnails) and the comments underneath the videos in the decision-making process using videos on quantum physics topics as an example. The study consists of two phases: In phase 1, N = 24 (pre-service) physics teachers were randomly assigned to one of three groups, each covering a different quantum topic (entanglement, quantum tunneling, or quantum computing, respectively). From eight options provided, they selected a suitable video for teaching while their eye movements were tracked using a stationary eye tracker in a laboratory setting, and think-aloud data was collected. In the subsequent phase 2, participants were allowed to freely choose one YouTube video on a second topic of the above-mentioned ones while thinking aloud. The results reveal a significant emphasis on video thumbnails during selection, with over one-third of the fixation time directed towards them. Think-aloud data confirms the importance of thumbnails in decision-making, e.g., as evidenced by a categorization of the study participants' arguments and thoughts voiced. A detailed analysis identifies that participants did not rely on (content-related) comments despite they have been found to be significantly correlated with the videos' explaining quality. Instead, decisions were influenced by surface features and pragmatic factors such as channel familiarity. Retrospective reflections through a questionnaire including rating scale items support these observations. Building on the existing empirical evidence, a decision tree is proposed to help teachers identify high-quality videos considering duration, likes, comments, and interactions. The decision tree can serve as a hypothesis for future research and needs to be evaluated in terms of how it can help systematize the process of selecting high-quality YouTube videos for science teaching.

Keywords Instructional videos · YouTube · Physics education · Quantum physics · Mixed-methods

✉ Philipp Bitzenbauer
philipp.bitzenbauer@fau.de

Extended author information available on the last page of the article

Introduction

Prior research has highlighted the benefits of YouTube for learning (e.g., see Jackman, 2019; Rosenthal, 2018) including increased engagement, better comprehension, and the flexibility to control the learning experience (Jebe et al., 2019; Kay, 2012; Stockwell et al., 2015). Consequently, educators and teachers have recognized the educational value of YouTube, employing it for instructional purposes (Jung & Lee, 2015; Manca & Ranieri, 2016). One of such purposes lies in the integration of YouTube explanatory videos into formal learning environments.

In the last years, there has been extensive science education research on explaining the quality of YouTube explanatory videos: Kulgemeyer (2020) developed a comprehensive framework for effective explanatory videos, based on guidelines published earlier in the literature (e.g., see Brame, 2016; Findeisen et al., 2019). Further studies have investigated the relationship between surface features, such as likes and views, and the explaining quality (i.e., the instructional quality) of YouTube explanatory videos (Kulgemeyer & Peters, 2016; Bitzenbauer et al., 2023): These studies have revealed that the surface features provided by YouTube may not serve as reliable indicators of the explaining quality of a specific video, while a statistically significant correlation was found between the number of content-related comments and the explaining quality. Based on the above findings, Bitzenbauer et al. (2023) emphasize that it is crucial to support teachers "in selecting videos with high explanation quality from the plethora of (online) resources" (p. 2) through evidence-based selection criteria.

However, to date, there is a dearth of studies investigating the video selection practices of physics pre-service teachers on YouTube, particularly concerning their decision-making factors. It remains unclear whether teachers rely on YouTube's provided metrics, such as likes, views, or the age of the video, or if they consider the comments section influential. This article addresses this research gap by presenting the findings of a mixed-methods study that explores the decision-making processes of (pre-service) physics teachers when selecting instructional videos on YouTube to be included in learning environments (see "Methods"). The study employs a combination of eye-tracking, think-aloud interviews, and a retrospective questionnaire survey to gain comprehensive insights into the thought processes and strategies employed by pre-service physics teachers during the video selection process.

Research Background

Explanatory videos, also referred to as instructional videos, play a vital role in science education research, for example, serving as concise introductions and explanations of specific topics of interest (Wolf & Kratzer, 2015). Explanatory videos typically do not exceed 10 min in length and have garnered increased attention in both formal and informal learning environments, especially through platforms such as YouTube (e.g., see Beautemps and Bresges, 2021; Pattier, 2021).

Quality Criteria of Instructional YouTube Videos

Recent scholarly investigations have focused on understanding the factors contributing to the success and popularity of explanatory YouTube videos, particularly in the field of science (Beautemps & Bresges, 2021; Welbourne & Grant, 2016). Notably, the video structure has emerged as a crucial determinant in this regard (Beautemps & Bresges, 2021).

However, the primary objective of explanatory videos is to support student learning, making the quality of explanations of utmost importance (Kulgemeyer & Wittwer, 2023; Pekdag & Le Marechal, 2010). Researchers have explored various frameworks and guidelines to enhance the effectiveness of explanatory videos. For example, Kulgemeyer (2020) proposed a comprehensive framework for creating effective explanatory videos that aligns with guidelines established earlier by Brame (2016) and Findeisen et al. (2019). Furthermore, this framework incorporates insights from multimedia learning research and draws upon studies related to instructional explanations conducted by Geelan (2012) and Wittwer and Renkl (2008). The framework encompasses seven factors comprising 14 features that collectively influence the effectiveness of explanatory videos. These factors include video structure, language-level adaptation, minimal digressions, consideration of prior knowledge, misconceptions, and student interest (Kulgemeyer, 2020).

Kulgemeyer and Wittwer (2023) empirically tested the effectiveness of the framework by comparing student achievement when exposed to videos developed in accordance with the framework against those that did not strictly adhere to the guidelines. The results demonstrated that students exposed to videos closely aligned with the framework exhibited significantly higher levels of declarative knowledge in post-tests ($d = 0.42$), although no statistically significant difference was observed in post-test scores related to conceptual knowledge.

The correlation between video metrics provided by YouTube, such as the number of views or likes, and the videos' explaining quality has yielded mixed results: Kulgemeyer and Peters (2016) conducted an exploratory study focusing on instructional YouTube videos on mechanics topics and found that the number of content-related comments posted by users below a specific video was the only variable that correlated significantly with the explaining quality. Conversely, the number of views, likes, and dislikes did not exhibit significant correlations. Similar findings have been brought forth by Kocyigit and Akaltun (2019) who evaluated 53 online videos using the the Global Quality Scale and found that the YouTube metrics did not significantly differ across quality groups. Bitzenbauer et al. (2023) conducted an additional exploratory study that specifically examined explanatory YouTube videos on quantum topics, namely quantum entanglement and quantum tunneling. In contrast to earlier findings, the authors observed a small but significant correlation between the number of likes and the quality of explanations in their sample of quantum topic videos ($r = 0.37$, $p < 0.01$).

Selection Processes of YouTube Explanatory Videos

The increasing abundance of low-quality educational content on YouTube has become a matter of concern for researchers (e.g., see Bohlin et al., 2017; Neumann and Herodotou, 2020; Tan, 2013) highlighting the crucial role that teachers play in selecting explanatory videos of high quality (Chtouki et al., 2012; Jones and Cuthrell, 2011). This issue is further exacerbated by the reliance on popularity-based rankings in search systems: For instance, Chelaru et al. (2012) observed that the top ten videos in the YouTube search results received a disproportionately higher number of views, likes, and comments. Additionally, the study by Chavira et al. (2021) revealed that out of the ten most-viewed videos analyzed in their study, only four were deemed satisfactory in terms of quality.

Despite existing research on YouTube video selection, to the best of our knowledge, no studies have specifically examined the process by which teachers select videos from the list provided by YouTube based on search queries. However, several studies have shed light on user behavior, indicating that individuals often sequentially view the returned videos until they find one that aligns with their needs (e.g., see Fyfield et al., 2021; Tan and Pearce, 2011).

The abundance of choices available on YouTube may contribute to choice overload, making it challenging to identify high-quality content (Toffler, 1984). Choice overload describes the phenomenon of increased difficulty in decision-making when faced with a large number of choices (Schwartz, 2016), potentially resulting in decreased motivation to engage with individual options (Iyengar & Lepper, 2000).

Against this backdrop, it becomes even more apparent that it is essential to support teachers in the process of selecting explanatory videos for classroom practice. Two main measures have been at the center of the debate so far:

1. Ranked lists of educational channels have been published to "help Internet users to narrow down their search space by recommending channels" (Tadbier and Shoufan, 2021, p. 3079). However, "there is no reason to assume that the extensive offer of ranked lists would not lead to choice overload" (Tadbier and Shoufan,2021, p. 3079).
2. To tackle these challenges, scholars have suggested the utilization of decision-assistance tools like meta-search engines, which employ aggregation techniques (Dwork et al., 2001; Haveliwala, 2002; Meng et al., 2002) as described by Tadbier and Shoufan (2021).

Research Rationale

While we agree that pre-made lists or similar resources might not optimally support (science) teachers in the process of selecting YouTube explanatory videos for classroom practice, we believe that the existing empirical evidence on the explaining quality of YouTube explanatory videos might indeed be useful to facilitate the systematization of teachers' decision-making process. As sketched in "Quality Criteria Of Instructional YouTube Videos," several studies have brought forth hints for instructional quality of online explanatory videos and might, hence, provide evidence in this regard:

- Bitzenbauer et al. (2023) found a statistically significant correlation ($r = 0.46$, $p < 0.001$) between explaining quality and the number of content-related comments in YouTube videos on quantum topics. Similarly, Kulgemeyer and Peters (2016) reported a significant correlation ($r = 0.38$, $p < 0.01$) between explaining quality and the number of relevant comments for videos on Newton's third law and Kepler's laws, respectively.
- Bitzenbauer et al. (2023) discovered a significant correlation ($r = 0.37$, $p < 0.01$) between the number of likes and video explaining quality.
- It is important to exercise caution when considering additional metrics provided by YouTube, such as the number of views, as previous research has not found stable correlations with explaining quality.

Of course, reviewing comments under each video in search of high-quality explanatory videos is practically unfeasible and time-consuming. Moreover, based on the available evidence, it is challenging to determine a quantitative threshold indicating an adequate number of content-related comments or likes. Nonetheless, considering the current state of research, it appears feasible to explore ideal (i.e., efficient) decision-making processes by systematically analyzing the order in which the different criteria can be employed by teachers: As a starting point for future research aimed at supporting teachers in their decision-making processes when selecting YouTube explanatory videos for science teaching, we propose the decision tree presented in Fig. 1 as a representation summarizing the hints of instructional quality of explanatory videos according to the current state of research described above. It is noteworthy that the proposed order is not to be considered strict.
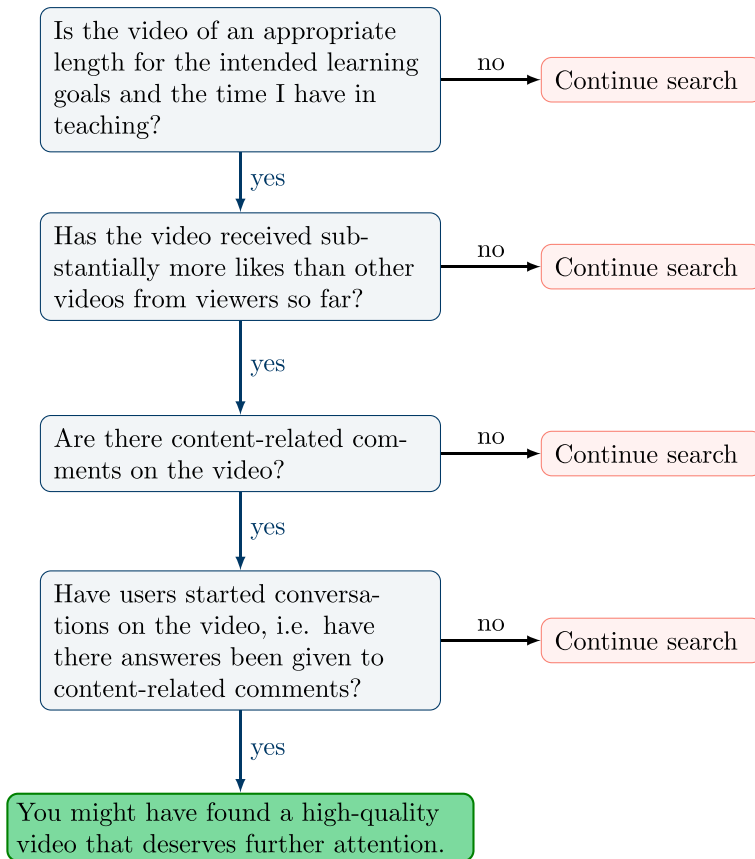
Fig. 1 Decision tree to support (pre-service) teachers selection process when searching explanatory videos suitable for physics teaching as hypothesized based on the current state of the literature

This decision tree suggests teachers to first ask quick initial questions to themselves during their search process such as whether a given video has an appropriate duration for classroom use or whether it has already received user likes. If both of these surface-level criteria are positive, it recommends teachers to explore the comments section. The presence of not only superficial but also content- or video-related comments indicates cognitive stimulation of viewers, as "videos that accumulate plenty of those relevant comments are more successful in catching viewers' attention as these videos might use either a more stimulating explanation or the explanation delivered is considered as a starting point for further learning progress" (Kulgemeyer & Peters, 2016, p. 12). Moreover, if there are even interactions among users, including responses to content-related comments, this may provide additional evidence of a high-quality video. Finally, teachers are then encouraged to assess the instructional quality of the video by viewing it themselves.

The suggested decision tree serves as a hypothesis for future studies examining teachers' selection processes of instructional videos for science teaching as stated above. The main objectives of this study are twofold: First, we aim to explore how pre-service teachers utilize YouTube metrics when selecting instructional videos for physics teaching. Secondly, we intend to compare their selection and decision-making processes with the procedure recommended by the existing literature, as represented through the decision tree in this article.

Hence, with our research, we aim to approach the following research question: How—if at all—do pre-service teachers use the features and comments sections provided by YouTube when selecting YouTube explanatory videos for teaching purposes?

We decided to address this question in the context of quantum physics YouTube videos because (a) as mentioned above, there have been related studies published previously we can build upon with our findings and (b) quantum physics deals with difficult-to-grasp topics and different visualizations or explanations are common to describe the same phenomena due to their abstract nature. Thus, a highly varying degree of explaining quality is to be expected when exploring explanatory videos for sub-topics of quantum physics like quantum tunneling or quantum entanglement.

Methods

Study Design

We investigated our research question by conducting a mixed-methods study comprised of eye-tracking, think-aloud interviews, and a concluding questionnaire. The selection of precisely these methods as well as their interrelation will be explained more thoroughly in the following subsections "Eye-Tracking"–"Questionnaire." The mixed-methods study consisted of three phases (P1 to P3; for an overview, see Fig. 2):

P1: In the first phase, the participants were presented with a pre-constructed image chart showing eight different YouTube video suggestions for a specific topic via the original surface features provided by YouTube (e.g., thumbnail, length, title, views, upload date, channel name, number of subscribers). As additional information, we added the corresponding number of likes the videos have already received (cf. Fig. 4). The participants were then asked to select one of the offered explanatory videos that they deemed suitable for use with learners without prior knowledge in physics teaching on this topic. The videos displayed in the chart, however, could not be opened or watched, and instead, the selection had to be based solely on the provided features. In addition to tracking the eye movements during the selection process (cf. "Eye-Tracking"), the participants were prompted to voice their thoughts at all times in the sense of a think-aloud interview (cf. "Think Aloud").

P2: In the second phase, the image chart was removed, and the participants were now allowed to freely explore a previously opened YouTube search tab concerned with a second specific quantum topic. Again, the task was to select one of the videos suggested by YouTube for a hypothetical physics classroom lesson covering the specific topic with learners without prior knowledge. In contrast to the first phase, the participants' eye movement was no longer tracked, but they were now also allowed to open and watch the videos as well as scroll through the comment section. This way, the selection process could place a more pronounced focus on content-related reasons to substantiate the rather superficially guided process in phase 1. Similar to phase 1, all thought processes had to be conveyed verbally at all times.

P3: After the initial combination of eye-tracking and think aloud, the study concluded with a questionnaire given to students in retrospective which asked the study participants to reflect on the importance of the different surface features provided by YouTube in their selection processes (cf. "Questionnaire").
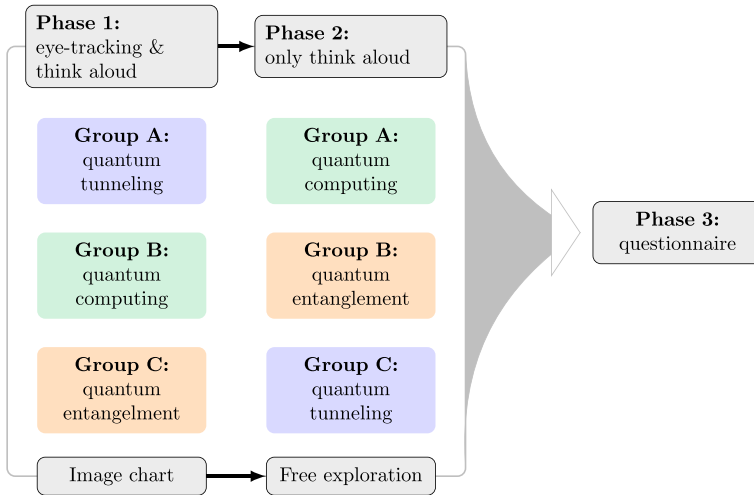
Fig. 2 Study design visualized using a flowchart. The different topics covered in the first two phases are color-coded and, as indicated, were switched among the groups A to C between both phases

To ensure that the results of our study are not directly dependent on (and hence, restricted to) a specific (quantum) topic covered throughout the phases, we randomly assigned each study participant to one of three different groups A, B, or C prior to starting the data collection. Each participant took part in the study individually and was then—depending on the group assignment—given the task of selecting explanatory YouTube videos on two different quantum topics (namely, either quantum tunneling, quantum entanglement, or quantum computing) in study phases 1 and 2 as described above. Kruskal-Wallis tests comparing each of the different eye-tracking metrics investigated in this study (cf. "Data Analysis") across the three study groups revealed no statistically significant differences among the groups (for all details on the test statistics, see the supplementary file). This indicates that our results are not directly linked to a specific topic, but it is sensible to analyze the data collected in this study across the groups—we took advantage of this observation in our study, as we analyzed the data from all 24 participants simultaneously, as if they had been collected under exactly the same conditions. An overview of our study design is presented in Fig. 2.

Data Collection

Eye-Tracking

Eye-tracking data was collected in study phase 1 using a stationary head-free eye-tracking system from Tobii (Tobii Pro Fusion) alongside their respective software (Tobii Pro Lab). The eye tracker operates at a sampling frequency of 60 Hz and a nominal spatial accuracy of $< 0.3°$ visual angle. The stimuli were presented on a 24-inch computer screen (1920×1080 pixel resolution and 60 Hz frame rate). Prior to the study, a nine-point calibration process was utilized to ensure accurate eye-tracking, and the participants were introduced to the basics of eye-tracking. The instructor verified the agreement between the measured gaze positions and the actual points on the screen. If the calibration results were not deemed satisfactory, the calibration was repeated. In instances where the eye tracker failed to detect sufficient

calibration data, participants were repositioned in front of the eye tracker. Additionally, potential factors that could have interfered with pupil detection were examined, e.g., lighting conditions, occlusion, or calibration drift. On average, the distance between each participant and the tracker was 60 cm.

Think Aloud

Since previous research has identified the need to complement eye-tracking data with additional verbal data, we amended the eye-tracking results by incorporating think-aloud interviews into our study design (Chien et al., 2015; Smith et al., 2010). During think-aloud interviews, "participants think out aloud while performing a given task, or recall thoughts immediately following completion of that task" (Eccles and Arsal, 2017, p. 514). The participants' verbalizations were recorded as well as transcribed and subsequently subjected to further analysis (cf. "Data Analysis"). The goal of this method lies in uncovering cognitive processes that are not as accessible with the other methods used (Rios et al., 2019). Thus, even though it might interfere with the study objective due to the verbalizations resulting in an overall higher cognitive demand, think-aloud studies are often used as an introspective annex (McKay, 2009; Sasaki, 2013). We leveraged this method by asking the participants to articulate their thought processes at any point in time in both, phases 1 (image chart) and 2 (free exploration). To use thinking aloud effectively, the participants were provided with an instruction on thinking aloud, which was formulated following Mackensen-Friedrichs (2004) to ensure a standardized procedure (Sandmann, 2014). Some of the cues given to the participants were (1) speak your thoughts aloud; (2) there should be no pauses in speaking, so verbalize your thoughts without pauses; (3) do not organize your thoughts before speaking, imagine you are alone in the room; and (4) thinking aloud may be a bit unfamiliar. Therefore, you will be supported and repeatedly prompted to express your thoughts.

The additional verbal data obtained from those interviews provides further insights into the cognitive processes, motivations, and decision-making that underlie the observed eye movements, offering a more complete picture of participants' experiences and interpretations (cf. Brückner et al., 2020). Furthermore, eye-tracking data alone can identify moments of attention shifts or fixations on specific elements, but it may not explain the reasons for these shifts. Since our study addresses selection processes based on visual stimuli, supplemental verbal data can clarify whether a shift in gaze was triggered by interest, confusion, or any other factors, shedding further light on the nature of the participants' attentional patterns.

Questionnaire

To further enhance cross-validity, we concluded our study with a final questionnaire in phase 3. Here, participants were presented with a list containing all (surface) features provided by YouTube and were asked to rate whether they considered the different features important to their decision-making processes on a four-point rating scale (strongly disagree, disagree, agree, strongly agree). The addressed features were number of views, likes, comments, and subscriptions as well as thumbnail, channel, video title, video length, video description, upload date, order determined by YouTube's search algorithm, and specific comments. On the one hand, these ratings allow to establish a ranking among all surface features in terms of their importance. On the other hand, the retrospective view obtained from the concluding questionnaire contrasts the introspective view from phases 1 and 2, enabling a triangulation with the findings from both the eye-tracking and the think-aloud interviews.

Sample

A total of $N = 24$ German pre-service physics teachers (9 female, 15 male) participated in this research. The participants were selected such that they (a) were at least in their second academic year and (b) did not rely on strong glasses or contact lenses (diopter > 1). Participation in our study was voluntary, not financially recompensed, and informed consent was obtained from all participants.

Data Analysis

The eye-tracking data were evaluated in terms of three metrics that reflect attention allocation and cognitive demand: First, we analyzed the *total fixation duration*, which can be described as "the total duration of all fixations on a specific stimulus" (van der Laan et al., 2015, p. 1). High values of this metric indicate a more pronounced focus on certain areas (Hahn & Klein, 2022); thus, it is the commonly reported measure of visual attention (Shruti Goyal & Miyapuram, 2015). Second, we investigated the metric *fixation counts* that often accompanies the total fixation duration as a measure of attention allocation (cf. Just and Carpenter, 1976; Wang et al., 2014). Lastly, we analyzed the *mean fixation duration*, which is often interpreted as a measure of cognitive processing demand (Negi & Mitra, 2020). Consequently, higher values of mean fixation duration indicate a "higher cognitive effort to process information" (Hahn and Klein, 2022, p. 10). The areas of interest (AOIs) required for quantitative analysis were matched with the surface features provided by YouTube (cf. "Eye-Tracking") as is indicated in Fig. 3. However, for the data analysis, the individual AOIs for each of the proposed videos shown in the image chart in phase 1 were not considered: Instead, so-called aggregated tags were created that combine the eye-tracking metrics for several related AOIs (e.g., all Like-AOIs). For example, the aggregated tag "Likes" was created, which summarizes all Like-AOIs for the individual video options, so that finally, for example, statements can
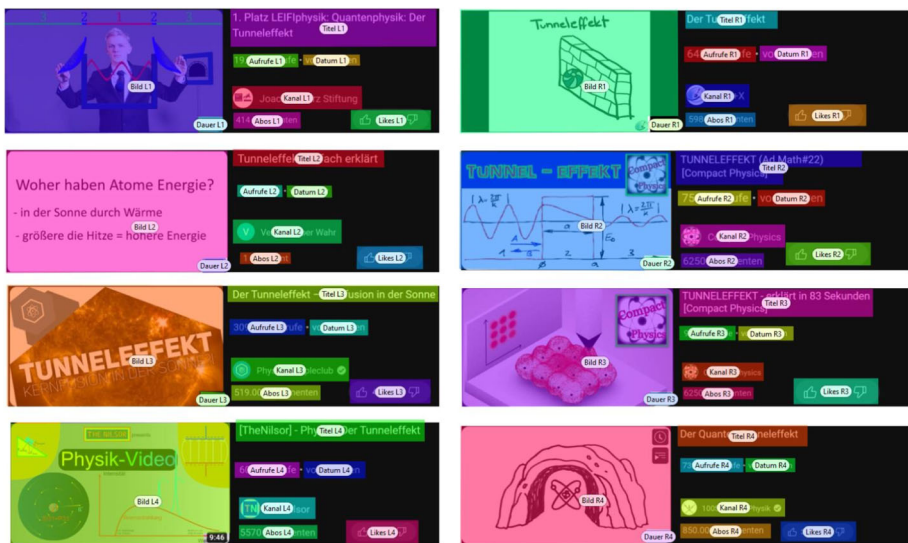

Fig. 3 The AOIs were defined covering all the surface features given for the eight video options shown to the participants. The eye-tracking metrics regarding the related AOIs were summarized using aggregated tags

be made about how often and how long the number of likes was viewed (across the different video options given).

For the subsequent think-aloud interviews, we conducted a qualitative content analysis. To this end, we (a) associated the participants' verbal expressions with the corresponding surface features provided by YouTube and (b) categorized their decisions for or against each video. The categories for these decisions were developed based on both inductive and deductive procedures (Forman & Damschroder, 2001). An overview of all categories and their descriptions is provided in the appendix (cf. Table 5). The category system was applied by two independent raters, and dissenting judgements were resolved through discussion. The interview data were analyzed threefold: First, we calculated the relative speaking time allocated to each (surface) feature and visually displayed the resulting share of each feature in a bar chart. Second, we visualized the temporal trajectory of each interview through the various categories and plotted them alongside a common axis, resulting in a temporal topography graphic for each of the two phases. Lastly, we counted the most frequently used arguments among the participants' reasonings and how often they lead to a decision for or against a video. This insight was used to provide an overview of the (surface) features provided by YouTube for each video that are most influential during the decision-making process of pre-service physics teachers.

The responses of the concluding questionnaire were summarized using a diverging stacked bar chart constructed from the participants' responses by aligning the bars from a stacked bar chart relative to the scale's center (Robbins & Heilberger, 2011). In addition, each of the response options was color-coded and equated with a number (strongly disagree $\hat{=} -2$, disagree $\hat{=} -1$, agree $\hat{=} 1$, agree completely $\hat{=} 2$) so that a mean agreement value for each surface feature could be calculated, resulting in a ranking among all surface features provided by YouTube (cf. Veith et al., 2022). It is important to note that rating scale data is ordinal in nature and as such the average results are merely being used as a quick means to represent the true data.

Results

In the following, we present the results of our study, separated by methodology. First, we provide an overview of the assessed eye-tracking metrics (cf. "Eye-Tracking Results"), and second, we enrich those findings with the results of the think-aloud data (cf. "Think-Aloud Interview Results") as well as the subsequent questionnaire study (cf. "Questionnaire Results").

Eye-Tracking Results

The eye-tracking data were collected in study phase 1 where participants were presented with a carefully constructed chart containing eight search results from YouTube (for details, see "Methods"). These options were specifically chosen to exhibit a range of surface features. The participants' task was to determine which of the eight explanatory videos would be suitable for inclusion in a learning environment related to the topic being investigated. Figure 4 presents an illustrative heat map generated from the eye-tracking data obtained from one of the participants in the study. The heat map provides visual information regarding the areas that received the highest attention during the task.

Table 1 provides the descriptive statistics for the metric *total fixation duration* for each area of interest. With a mean percentage of 35.39% of the total fixation duration, the thumbnail

Fig. 4 Exemplary heat map created from the eye-tracking data from one of the study participants in phase 1 of our study. Qualitatively, participants' spots of attention are represented through red color

was by far the most compelling AOI and the only one with a share above 10%. While the AOIs' title (9.58%) and channel (6.17%) were also able to captivate the participants' attention to some extent, the remaining AOIs played a seemingly negligible role during the selection process. This discrepancy is visualized via boxplots in Fig. 5.

Descriptive statistics for the metric *mean fixation duration* are provided in Table 2. The data in this regard paint a different picture by taking similar values for each AOI. The mean fixation durations average between 176 (subs) and 240 ms (date) and, thus, lie in the typical range of 100–600 ms reported by Hahn and Klein (2022). The boxplots illustrating these data substantiate this relationship—the middle 75% of data can be located between 200 and 300 ms across almost all AOIs (cf. Fig. 6).

Lastly, we investigated the metric *fixation counts* for each AOI. With there being striking differences in the total fixation durations of each AOI but similar values for mean fixation duration, it becomes apparent that the number of fixations must vary in a manner similar to the total fixation duration. The data provided in Table 3 paint a coherent picture in this

Table 1 Descriptive statistics for the total fixation duration (in percent) for each area of interest

	Subs	Views	Thumbnail	Date	Length	Channel	Likes	Title
Max	6.48	5.48	57.03	4.85	3.66	14.53	5.24	19.10
Min	0.19	0.57	12.38	0.00	0.00	1.47	0.00	2.39
Q_1	1.28	1.73	29.84	0.29	0.46	4.73	0.32	6.69
Q_2	1.89	2.20	37.39	0.56	0.71	5.90	0.95	9.45
Q_3	3.61	3.72	41.73	0.94	1.50	7.84	1.79	12.39
Mean	2.55	2.47	35.39	0.78	1.08	6.17	1.56	9.58

In addition to the minimum, maximum, and mean values, we report the lower (Q_1), middle (Q_2), and upper (Q_3) quartiles
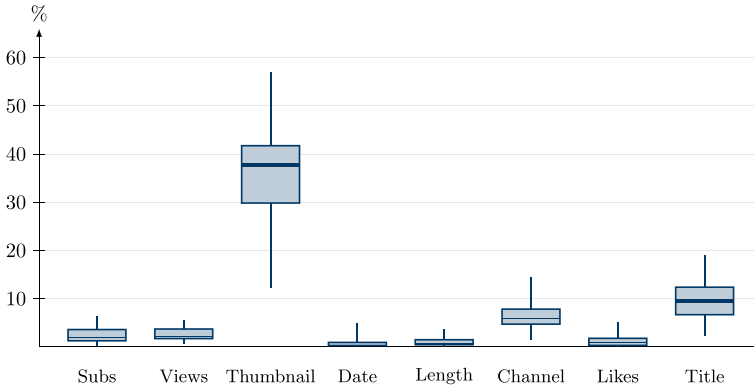
Fig. 5 Boxplots for the total fixation duration for each area of interest. The whiskers indicate $1.5 \times$ IQR, where IQR is the interquartile range

Table 2 Descriptive statistics for the mean fixation duration (in seconds) for each area of interest

	Subs	Views	Thumbnail	Date	Length	Channel	Likes	Title
Max	0.31	0.41	0.31	0.31	0.44	0.35	0.37	0.32
Min	0.16	0.14	0.16	0.08	0.09	0.13	0.13	0.13
Q_1	0.21	0.20	0.21	0.18	0.20	0.24	0.17	0.23
Q_2	0.23	0.23	0.25	0.22	0.23	0.26	0.21	0.25
Q_3	0.27	0.26	0.26	0.26	0.31	0.28	0.27	0.27
Mean	0.17	0.19	0.21	0.24	0.18	0.23	0.21	0.23

In addition to the minimum, maximum, and mean values, we report the lower (Q_1), middle (Q_2), and upper (Q_3) quartiles
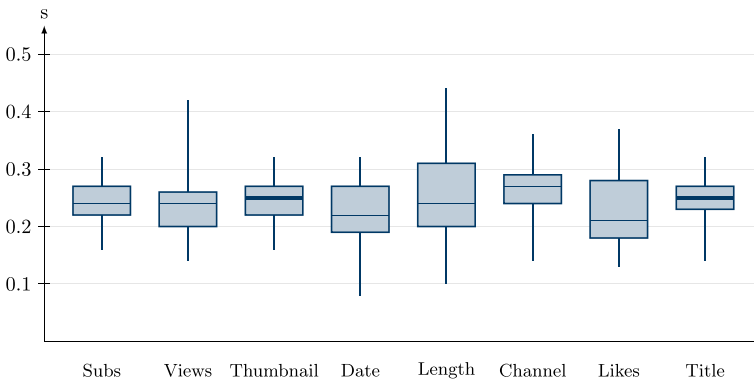

Fig. 6 Boxplots for the mean fixation duration for each area of interest. The whiskers indicate $1.5 \times$ IQR, where IQR is the interquartile range

Table 3 Descriptive statistics for the fixation count (in %) for each area of interest

	Subs	Views	Thumbnail	Date	Length	Channel	Likes	Title
Max	6.50	5.34	69.30	5.20	2.93	14.93	5.60	22.63
Min	0.37	0.94	18.34	0.00	0.00	2.60	0.00	5.14
Q_1	1.83	2.14	36.39	0.46	0.59	5.25	0.64	8.69
Q_2	2.29	3.04	47.29	0.93	1.04	6.55	1.00	11.29
Q_3	4.67	4.21	50.84	1.25	1.59	8.41	2.83	14.96
Mean	3.04	3.11	45.59	1.06	1.12	7.12	1.82	11.60

In addition to the minimum, maximum, and mean values, we report the lower (Q_1), middle (Q_2), and upper (Q_3) quartiles

regard. Again, with a mean percentage of 45.59% of the total number of fixations, there is a predominant focus on the AOI thumbnail, with title (11.60%) and channel (7.12%) being second and third, respectively. Consequently, the boxplots for the metrics total fixation duration and fixation count are somewhat congruent (cf. Fig. 7).

Think-Aloud Interview Results

To analyze the selection process more thoroughly, the results from the eye-tracking study are now complemented with data from a think-aloud study, as described in "Methods." Figure 8 offers an initial comprehensive overview of the content aspects addressed in the argumentation provided by the study participants: It shows the relative proportion (of total speaking time) of each surface feature in participants' utterances in both study phases. In phase 1, the most pronounced focus was placed on the thumbnail with 30.9%, followed by channel and title with 23.0% and 14.2%, respectively. Regarding the free exploration in phase 2, however, the data convey a more differentiated impression: Here, the surface feature channel emerges on the top with 22.0%, with the thumbnail (20.0%) and length (17.4%) taking second and third places. In addition, the surface features likes, subs, date, order, description, and comments are almost negligible with an allocated speaking time of below 5% throughout both phases.

A more comprehensive insight into the structure of the participants' selection process is offered by the bar charts in Figs. 9 and 10.
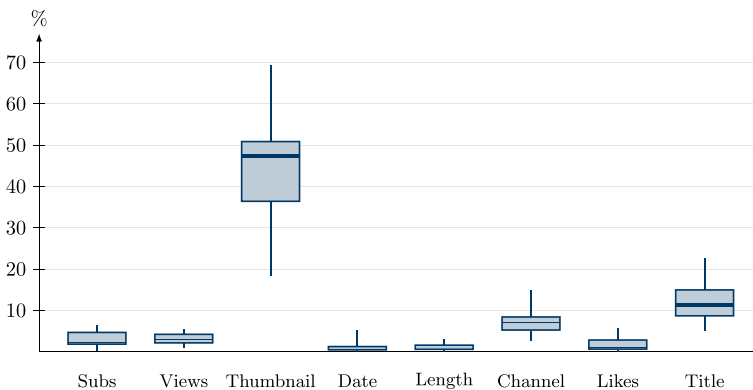

Fig. 7 Boxplots for the fixation counts for each area of interest. The whiskers indicate $1.5 \times$ IQR, where IQR is the interquartile range
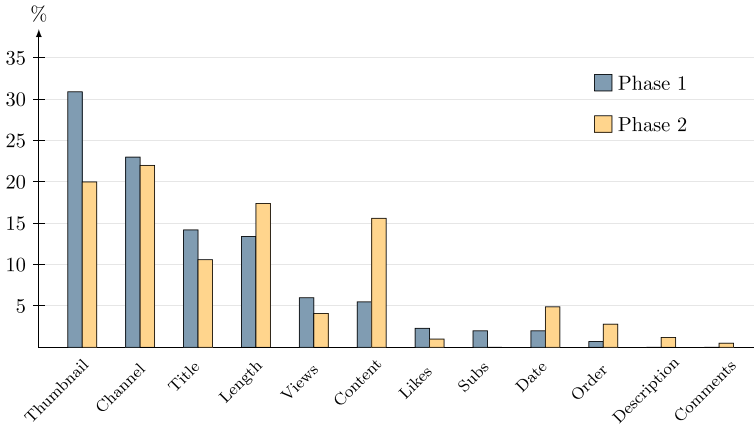
Fig. 8 Bar chart visualizing the relative proportion (of total speaking time) of each surface feature in participants' utterances in both phases. Phase 1 indicates the allocated speaking time regarding the pre-constructed image chart, while phase 2 indicates the allocated speaking time during free exploration (cf. "Methods")

Here, each individual interview is presented as a bar, and the sections dedicated to the different surface features are color-coded respectively. Hence, these visualizations allow for a deeper insight into the temporal topography of each interview. Analyzing this topography, it becomes apparent that blue (thumbnail) and violet (channel) cover the most area during the first phase, in accordance with the findings presented in Fig. 8. In the second phase, where participants were allowed to click on and even watch videos, this dynamic changes: On the one hand, the violet sections increase, indicating a greater focus on the surface feature
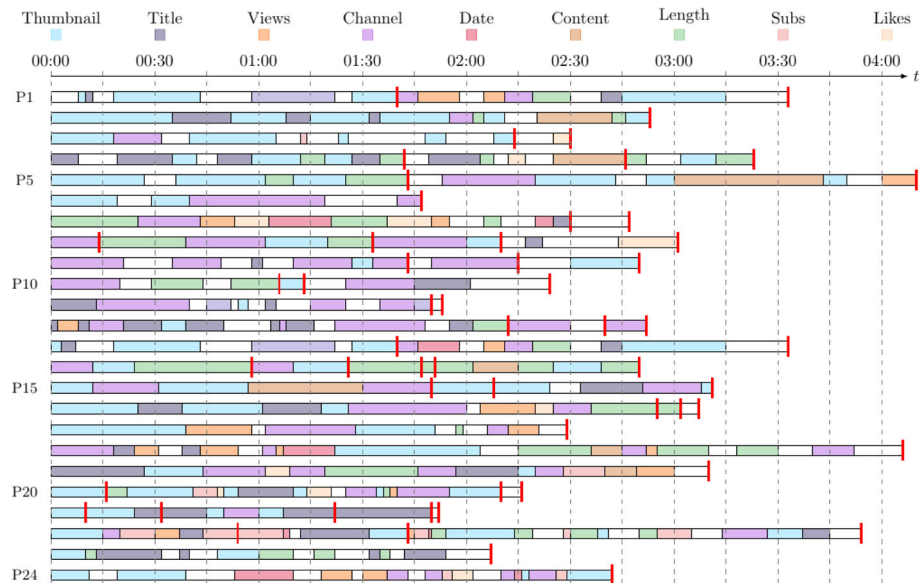

Fig. 9 Topography of the think-aloud interviews in the first phase, one for each study participant P1 to P24. The red strokes indicate a decision for or against a video. The upper row indicates the color coding for each surface feature. White sections represent small breaks where the participants did not address a specific surface feature
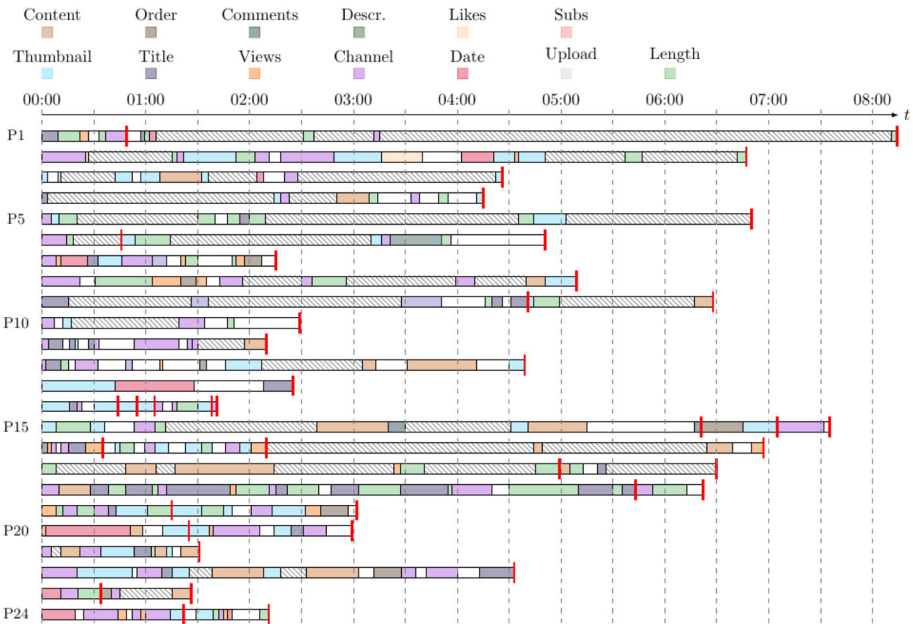
Fig. 10 Topography of the think-aloud interviews in the second phase, one for each study participant P1 to P24. The red strokes indicate a decision for or against a video. The upper row indicates the color coding for each surface feature. Hatched sections represent parts of the interview where the participants watched a video

channel. On the other hand, the participants had access to more surface features such as comments or the video description. Decisions for or against a video are indicated by red strokes. For example, a red stroke after a blue section displays a decision against a video because of the thumbnail. A summary of all decisions and the surface features they were based on is presented in Table 4.

With a total of 26 and 23 decisions, the surface features thumbnail and video length are by far the most influential ones. In congruence to the bar chart presented in Fig. 8, the channel, the video title, and the number of views can also be regarded as guides for the selection process, while more specific features such as likes, subs, or comments seem almost entirely irrelevant for decision-making. Lastly, it is noticeable that most decisions could not be attributed to specific surface features. In particular, 30 out of the 51 positive decisions do not seem to be related to surface features provided by YouTube. We will elaborate on this finding in more detail in "Discussion."

To obtain a more in-depth view on the decisions used by participants that could be related to a surface feature, their arguments during the interviews were categorized (cf. "Methods"). Since the thumbnail feature leads to the most decisions in both phases, we present a bar chart for the most frequently used arguments in Fig. 11.

With the most frequently used arguments being T3 (the thumbnail indicates an interesting video) and T4 (the thumbnail indicates a boring or nonprofessional video), it becomes apparent that arguments addressing affective aspects dominate over content-related reasonings. An overview of the overall top five most frequently used arguments is provided in Fig. 12. We discuss these findings in "Discussion."

Table 4 Overview of the decisions for or against a video based on the respective (surface) features, sorted by study phases

Surface feature	Decision for (+) and against (−) a video in phase 1		Decision for (+) and against (−) a video in phase 2		Total decisions for both phases combined
	+	−	+	−	
Thumbnail	3	12	2	9	26 (5+, 21−)
Length	3	12	2	6	23 (5+, 18−)
Channel	3	4	0	2	9 (3+, 6−)
Title	1	4	1	3	9 (2+, 7−)
Views	1	1	1	3	6 (2+, 4−)
Content	0	1	2	0	3 (2+, 1−)
Likes	1	1	0	0	2 (1+, 1−)
Subs	0	1	0	0	1 (0+, 1−)
Description	0	0	1	0	1(1+, 0−)
Order	0	0	0	0	0
Date	0	0	0	0	0
Comments	0	0	0	0	0
Not related to a specific surface feature	13	1	17	0	31 (30+, 1−)

Questionnaire Results

In the final part of our study we, in retrospection, asked the participants to evaluate the extent to which they agree with the respective features having been relevant for their selection
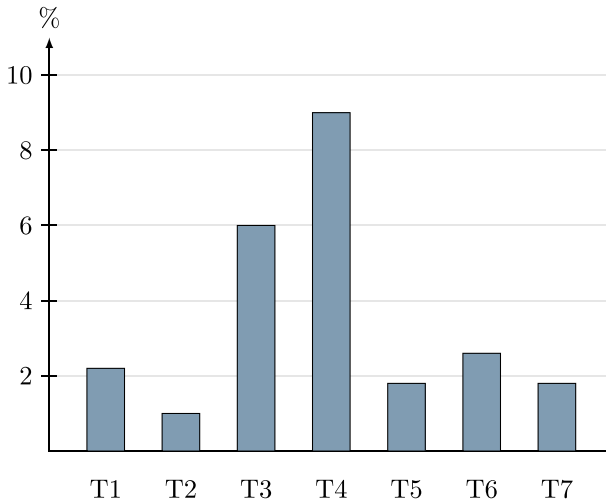

Fig. 11 Bar charts visualizing the relative share of each category (T1 to T5) addressing the thumbnail that was used as an argument for or against a video. The respective category system is provided in the appendix (cf. Table 5)
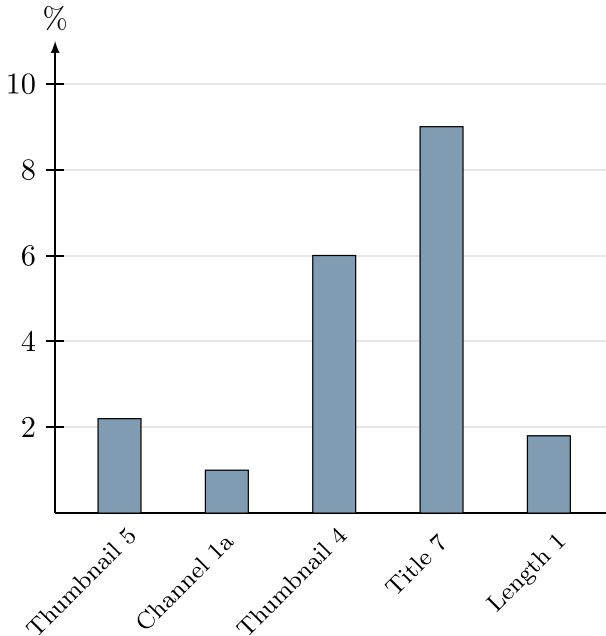
Fig. 12 Top five of the most frequently used arguments for or against a video during the think-aloud interviews. The respective category system is provided in the appendix (cf. Table 5)

process. The responses are visualized using a diverging stacked bar chart in Fig. 13. In congruence to our previous findings, the thumbnail takes a sole first place with a rating of 1.80 (where 2 corresponds to "agree completely" and −2 to "strongly disagree"). With 21 out of 24 participants agreeing completely with the thumbnail being important for their selection process, the thumbnail even exceeds the video length in terms of relevance for

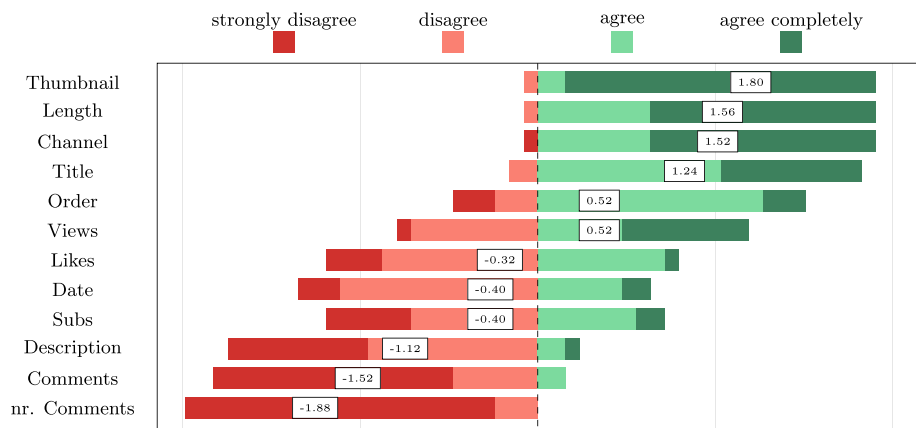**Fig. 13** Diverging stacked bar chart visualizing the participants' agreement that the respective surface feature is important for decision-making. The respective average ratings for each (surface) feature (cf. "Methods") are provided in labels on each bar where 2 corresponds to "agree completely" and −2 to "strongly disagree." The abbreviation "nr. Comments" stands for the number of comments under a video

decision-making in the participants' retrospective views. In contrast, the number of comments (average rating of -1.88) and comments themselves (average rating of -1.52) solidify last places and do not seem to contribute meaningfully to the decision-making process.

Discussion

The Process of Selecting Instructional Videos for Physics Teaching

In phase 1 of our study, pre-service physics teachers were given the task of selecting one explanatory video on quantum physics from a set of eight options. Participants were provided with excerpts from the YouTube search results and had access to various metrics such as views, likes, and channel information. The analysis of participants' eye movements revealed a significant emphasis on video thumbnails, with more than one-third of their total fixation time and counts directed towards this area of interest (AOI). Surprisingly, there were no statistically significant differences in mean fixation duration between different AOIs, despite this measure "is often considered an indicator of cognitive processing demand" (Hahn and Klein, 2022, p. 10). This finding, hence, contrasts with the results of Hsieh and Chen (2011), who suggested that viewing content with different information types requires varying cognitive resources.

To gain further insight into the participants' selection processes, we examined their think-aloud data. This analysis, consistent with the eye-tracking data, revealed that during both phase 1 (selection of one out of eight options based on surface features) and phase 2 (free exploration), participants predominantly voiced their considerations in relation to the AOI thumbnail. An in-depth categorization of argumentation structures and decision-making uncovered four key observations:

1. The video duration played a significant role in participants' choices, aligning with didactic perspectives as they were selecting videos for instructional purposes.
2. The video content had a minor influence on participants' decisions during the free exploration phase. Instead, choices were primarily guided by thumbnail, duration, channel, and title features, indicating a reliance on surface features and pragmatic decision-making among pre-service physics teachers. This tendency to select videos they already had a connection with or were familiar with, such as those from known channels, is in accordance with findings from cognitive psychology (Chen, 2016).
3. A notable portion of the decisions made during the study could not be attributed to surface features, comments, or video content based on either eye movements or verbalizations. Hence, in these cases, the study participants either struggled to explicitly articulate their decisions due to multiple considerations or did not just articulate them at a deeper level. Similar cases have been observed in physics education research on teachers' professional competences, where prior research has found that experienced teachers' actions in the classroom are guided by informed decisions and teaching routines that cannot be easily verbalized (e.g., see Borko and Livingston, 1989; Livingston and Borko, 1989; Stender, 2014). To clarify whether similar principles contribute to an explanation of the observations made in our study requires further investigation.
4. Although the participants had the opportunity to view comments associated with each YouTube video during the free exploration phase 2, surprisingly, none of the participants explicitly based their decisions on comments. This observation contrasts with findings from previous research reporting that students relied on comments as an indicator of video quality (e.g., see Fyfield et al., 2021; Tan and Pearce, 2011) and, instead, indicates that the selection process tends to be less systematic. However, previous studies have consistently

shown a strong and statistically significant correlation between the explanatory quality of YouTube videos and the number of content-related comments (Kulgemeyer & Peters, 2016; Bitzenbauer et al., 2023). It is therefore noteworthy that these comments did not play a significant role in the decision-making process of our participants.

The eye-tracking and think-aloud data were complemented by retrospective questionnaire responses: When asked about the features that influenced their video selection for instructional purposes, the majority of respondents indicated thumbnail, duration, channel, or title, while video descriptions and the quantity or quality of comments played a minor role, even in retrospective evaluation.

The finding that participants primarily explored the top results of the YouTube search list aligns with previous studies (e.g., see Fyfield et al., 2021; Tan and Pearce, 2011): Over half of the participants reported approaching video selection in a sequential manner based on the order of videos in the search list (cf. Fig. 13). This pragmatic approach leads to quick decisions (made in a time frame of less than 10 min in the free exploration phase 2 of this study) that are mainly based on surface features (e.g., thumbnails) or familiarity (e.g., channel). The analysis of individual argumentation categories (see Figs. 11 and 12) supports this assumption.

In conclusion, our findings suggest that the decision-making process of (pre-service) physics teachers when searching for suitable YouTube explanatory videos (on quantum topics in this study) for instructional purposes is primarily driven by pragmatism, efficiency, and reliance on familiar features. The availability of empirical evidence regarding the explanatory quality of online videos seems to be overlooked by (pre-service) physics teachers, representing a missed opportunity to streamline the selection process. Therefore, in "Conclusion," we will synthesize existing empirical evidence and propose a preliminary decision tree that may assist teachers in efficiently identifying high-quality explanatory videos on YouTube.

Contrasting the Selection Processes with the Proposed Decision Tree

The observations made in this study indicate that the selection processes of (pre-service) physics teachers when searching explanatory videos suitable for physics teaching are predominantly unsystematic, relying on superficial or familiar aspects, and characterized by pragmatic choices. These tendencies give rise to two interconnected issues when it comes to real instructional preparation, where videos with high explaining quality are sought:

1. Teachers may require significant time to find suitable videos due to the unsystematic approach.
2. There is a probability that teachers may select videos of lower quality.

It is obvious that in the study reported in this article, the decision-making process of the participants in most cases diverged from the proposed decision tree. In light of this, it becomes necessary to support teachers in systematizing their selection process to overcome the identified problems in practice. The decision tree proposed in "Research Rationale" might be a valuable tool in this regard as it reflects the state of the literature. While we are aware that future studies are required and might lead to a refined version of the decision tree, the significance of the decision tree in its current form lies in its capacity to systematize the selection process without imposing quantitative guidelines or thresholds. This acknowledges the absence of empirical evidence supporting such measures and recognizes that decisions should be made by teachers on a case-by-case basis, taking into account the specific topic. Future studies should investigate whether the use of the decision tree indeed facilitates efficient and

successful identification of high-quality explanatory videos on various topics. Additionally, it will be crucial to determine whether decision steps need to be supplemented or specified.

Limitations

The present study has several limitations that should be considered when interpreting the results. First, the focus on explanatory videos related to three specific quantum topics may restrict the generalizability of the findings. Although we designed the study with three separate groups, each tasked with selecting videos for instructional situations on two different quantum topics, this control measure may not account for potential variations that could arise if explanatory videos on further (physics) topics, e.g., classical mechanics topics, were included. Further research is needed to validate the reported results in a broader range of topics. Second, understanding the cognitive processes of pre-service physics teachers during video selection is a complex empirical endeavor, and the chosen data collection methods—even though they might complement each other—come with inherent limitations. While the analysis of eye-tracking data is based on the eye-mind assumption (Just & Carpenter, 1980), previous research has emphasized the importance of complementing eye movement analysis with additional verbal data to gain a comprehensive understanding (Brückner et al., 2020; Chien et al., 2015; Chiou et al., 2022; Mason et al., 2013; Smith et al., 2010; Wu & Liu, 2021). To address this concern within our mixed-methods approach, we employed introspective thinking-aloud in our study. Additionally, the retrospective questionnaire used for internal validation allows participants to reflect on their experiences; however, it may also trigger ad hoc generated associations and thoughts regarding the different YouTube surface features (for similar arguments, see Winkler et al., 2021; Winkler et al. 2023). Third, it is important to consider that while this study focuses on the selection processes employed by (pre-service) physics teachers in finding YouTube explanatory videos on quantum physics suitable for teaching, the selection situations created within the study design differ from real classroom planning scenarios. Particularly, in our study, participants had unlimited time for decision-making, whereas real instructional planning is significantly influenced by time constraints. However, the analysis of think-aloud data reveals that decision-making occurred within a time frame of less than 10 min in phase 2 of the study (free exploration), which aligns with a reasonable time frame in natural classroom planning situations. Lastly, the time-consuming nature of the study led to a relatively small sample size. However, the primary aim of this study was to gain in-depth insights into the video selection process rather than to produce generalizable findings on a surface level. Future research with larger sample sizes could provide a broader perspective on the topic.

Conclusion

This mixed-methods study explored how pre-service teachers select instructional videos on YouTube for physics teaching, focusing on the role of surface features (likes, views, thumbnails) and comments. The results indicate that the decision-making processes of (pre-service) physics teachers when searching for suitable YouTube explanatory videos are primarily driven by pragmatism, efficiency, and reliance on familiar features.

Based on the current state of research into the explaining quality of online explanatory videos, we proposed a decision tree which reflects how an efficient and successful selection process might look. Although the decision-making process of the study participants often

differed from the proposed decision tree in this study, it serves as a hypothesis for future research aimed at supporting teachers in systematizing their selection process: Further studies should explore whether the decision tree facilitates efficient and successful identification of high-quality explanatory videos on various topics and how it might be adapted and refined, e.g., with regard to different subject areas and teaching contexts. Also, future studies might examine how the decision tree works as a tool for preparing teaching, e.g., in related courses in science teacher education. Lastly, it seems particularly crucial to consider the evolving nature of online platforms in future research: For example, research could examine how the decision tree (or an evolved version thereof) can be adapted to accommodate changes in platform features and the emergence of new video metrics. Collaborative research involving educators, researchers, and platform developers may further enhance the decision tree's practicality and usability for (pre-service) teachers, facilitating their video selection process and ultimately benefiting student learning experiences in physics and beyond.

Appendix

Table 5 Category system used for the think aloud study

Surface feature	Category
Thumbnail	1. The thumbnail indicates a thematically well-suited video
	2. The thumbnail indicates that the video might be off-topic
	3. The thumbnail indicates that the video might not be exhaustive, e.g., it is part of a longer series
	4. The thumbnail seems interesting
	5. The thumbnail seems unappealing or non-professional
	6. The thumbnail indicates that the topic is explained well
	7. The thumbnail indicates that the topic is not explained well
	8. Neutral examination of the thumbnail
Title	1. The title indicates a thematically well-suited video
	2. The title indicates that the video might be off-topic
	3. The thumbnail indicates that the video might not be exhaustive, e.g., it is part of a longer series
	4. The title seems interesting
	5. The title seems unappealing or non-professional
	6. The title indicates that the video might delve too deeply into the topic
	7. Neutral examination of the title
Subscriptions	1. The respective channel has many subscriptions
	2. The respective channel has few subscriptions
	3. The respective channel has more subscriptions than other channels
	4. The respective channel has fewer subscriptions than other channels
Channel	1. The channel is known
	(a) The channel's name is known

Table 5 continued

Surface feature	Category
	(b) The channel is known for high-quality videos
	(c) The channel is known for scientifically inaccurate videos
	(d) The channel is known for low-quality videos
	2. The channel is unknown
	3. Neutral examination of the channel
	4. The channel seems fitting
	5. The channel seems unfitting
	6. The channel is verified (blue checkmark)
Likes	1. The video has many likes
	2. The video has a few likes
	3. The video has more likes than other videos
	4. The video has fewer likes than other videos
Views	1. The video has many views
	2. The video has few views
	3. The video has more views than other videos
	4. The video has fewer views than other videos
	5. Neutral examination of views
Length	1. The video length seems too long
	2. The video length seems accurate
	3. The video length seems rather short
	4. The video length seems too short
	5. Netrual examination of video length
Upload	1. The video is too old
	2. The video is new
	3. Neutral examination of upload date
Order	1. The videos suggested first are fitting
	2. The videos suggested first are unfitting
Description	1. Neutral examination of video description
	2. The video description is appealing
Content	1. The video content could be fitting
	2. The video content seems rather unfitting
	3. Neutral examination of video content
	4. The video content is contingent
Comments	1. Checking the feedback emerging from comments
	2. The comments are rather positive

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The data presented in this study are available on request from the corresponding author.

Declarations

Ethical Approval Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements.

Informed Consent Informed consent was obtained from all subjects involved in the study to publish this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Beautemps, J., & Bresges, A. (2021). What comprises a successful educational science YouTube video? A five-thousand user survey on viewing behaviors and self-perceived importance of various variables controlled by content creators. *Frontiers in Communication, 5*, 600595.

Bitzenbauer, P., Höfler, S., Veith, J., Winkler, B., Zenger, T., & Kulgemeyer, C. (2023). Exploring the relationship between surface features and explaining quality of youtube explanatory videos. *International Journal of Science and Mathematics Education*, 1–24. Retrieved from https://doi.org/10.1007/s10763-022-10351-w

Bohlin, G., Göransson, A., Höst, G. E., & Tibell, L. A. (2017). A conceptual characterization of online videos explaining natural selection. *Science & Education, 26*, 975–999.

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American educational research journal, 26*(4), 473–498.

Brame, C.J. (2016). Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education, 15*(4), es6.

Brückner, S., Schneider, J., Zlatkin-Troitschanskaia, O., & Drachsler, H. (2020). Epistemic network analyses of economics students' graph understanding: An eye-tracking study. *Sensors, 20*(23), 6908.

Chavira, A.M., Rivas, J.F., Molina, A.P.R.F., de la Cruz, S.A., Zárate, A.C., Musa, A.B., & Osorio, V.J.C. (2021). The educational quality of the critical view of safety in videos on YouTube® versus specialized platforms: Which is better? critical view of safety in virtual resources. *Surgical Endoscopy, 1*–9

Chelaru, S.V., Orellana-Rodriguez, C., & Altingovde, I.S. (2012). Can social features help learning to rank YouTube videos? *International Conference on Web Information Systems Engineering* (pp. 552–566).

Chen, L. (2016). The things attracting our attention: Evidence from text reading. Advances in Cognitive Neurodynamics (v) *Proceedings of the Fifth International Conference on Cognitive Neurodynamics-2015* (pp. 573–578).

Chien, K.-P., Tsai, C.-Y., Chen, H.-L., Chang, W.-H., & Chen, S. (2015). Learning differences and eye fixation patterns in virtual and physical science laboratories. *Computers & Education, 82*, 191–201.

Chiou, G.-L., Hsu, C.-Y., & Tsai, M.-J. (2022). Exploring how students interact with guidance in a physics simulation: Evidence from eye-movement and log data analyses. *Interactive Learning Environments, 30*(3), 484–497.

Chtouki, Y., Harroud, H., Khalidi, M., & Bennani, S. (2012). The impact of YouTube videos on the student's learning. *2012 International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1–4)

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. Proceedings of the 10th International Conference on World Wide Web (pp. 613–622).

Eccles, D. W., & Arsal, G. (2017). The think aloud method: What is it and how do I use it? *Qualitative Research in Sport, Exercise and Health, 9*, 514–531.

Findeisen, S., Horn, S., & Seifried, J. (2019). Lernen durch Videos: Empirische Befunde zur Gestaltung von Erklärvideos. *MedienPädagogik,* 16–36.

Forman, J., & Damschroder, L. (2001). Qualitative content analysis. L. Jacoby & L. A. Siminoff (Eds.), *Empirical methods for bioethics: A primer* (pp. 39–62). Bingley, UK: Emerald Group Publishing Limited

Fyfield, M., Henderson, M., & Phillips, M. (2021). Navigating four billion videos: Teacher search strategies and the YouTube algorithm. *Learning, Media and Technology, 46*(1), 47–59.

Geelan, D. (2012). Teacher explanations. *Second International Handbook of Science Education*, 987–999.

Hahn, L., & Klein, P. (2022). Eye tracking in physics education research: A systematic literature review. *Physical Review Physics Education Research, 18*(1), 013102.

Haveliwala, T.H. (2002). Topic-sensitive pagerank. *Proceedings of the 11th international conference on world wide web*(pp. 517–526).

Hsieh, Y.-C., & Chen, K.-H. (2011). How different information types affect viewer's attention on internet advertising. *Computers in human Behavior, 27*(2), 935–945.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995.

Jackman, W. M. (2019). YouTube usage in the university classroom: An argument for its pedagogical benefits. *International Journal of Emerging Technologies in Learning (Online), 14*(9), 157.

Jebe, F., Konietzko, S., Lichtschlag, M., & Liebau, E. (2019). Study: "Youth/YouTube/cultural education. Horizon 2019" '. Essen, Germany: Rat für Kulturelle Bildung e.V.

Jones, T., & Cuthrell, K. (2011). YouTube: Educational potentials and pitfalls. *Computers in the Schools, 28*(1), 75–85.

Jung, I., & Lee, Y. (2015). YouTube acceptance by university educators and students: A cross-cultural perspective. *Innovations in Education and Teaching International, 52*(3), 243–253.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8*, 441–480.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review, 87*(4), 329.

Kay, R. H. (2012). Exploring the use of video podcasts in education: A comprehensive review of the literature. *Computers in Human Behavior, 28*(3), 820–831.

Kocyigit, B. F., & Akaltun, M. S. (2019). Does YouTube provide high quality information? Assessment of secukinumab videos. *Rheumatology international, 39*(7), 1263–1268.

Kulgemeyer, C. (2020). A framework of effective science explanation videos informed by criteria for instructional explanations. *Research in Science Education, 50*(6), 2441–2462.

Kulgemeyer, C., & Peters, C.H. (2016). Exploring the explaining quality of physics online explanatory videos. *European Journal of Physics*, *37*(6), 065705.

Kulgemeyer, C., & Wittwer, J. (2023). Misconceptions in physics explainer videos and the illusion of understanding: An experimental study. *International Journal of Science and Mathematics Education*, *21*(2), 417–437.

Livingston, C., & Borko, H. (1989). Expert-novice differences in teaching: A cognitive analysis and implications for teacher education. *Journal of teacher education, 40*(4), 36–42.

Mackensen-Friedrichs, I. (2004). *Promoting the acquisition of expertise through learning with example tasks in year 9 biology lessons*. University Kiel

Manca, S., & Ranieri, M. (2016). Facebook and the others. Potentials and obstacles of social media for teaching in higher education. *Computers & Education, 95*, 216–230.

Mason, L., Pluchino, P., Tornatora, M. C., & Ariasi, N. (2013). An eyetracking study of learning from science text with concrete and abstract illustrations. *The Journal of Experimental Education, 81*(3), 356–384.

McKay, S.L. (2009). Introspective techniques. J. Heigham & R.A. Croker (Eds.), Qualitative research in applied linguistics. Springer.

Meng, W., Yu, C., & Liu, K.-L. (2002). Building efficient and effective metasearch engines. *ACM Computing Surveys (CSUR), 34*(1), 48–89.

Negi, S., & Mitra, R. (2020). Fixation duration and the learning process: An eye tracking study with subtitled videos. *Journal of Eye Movement Research,* 13.

Neumann, M. M., & Herodotou, C. (2020). Evaluating YouTube videos for young children. *Education and Information Technologies, 25*, 4459–4475.

Pattier, D. (2021). Science on YouTube: Successful edutubers. TECHNO REVIEW. International Technology, Science and Society Review/Revista Internacional de Tecnología, Ciencia y Sociedad, 10(1), 1–15.

Pekdag, B., & Le Marechal, J. F. (2010). Movies in chemistry education. *Asia-Pacific Forum on Science Learning and Teaching, 11*(1), 15.

Rios, L., Pollard, B., Dounas-Frazer, D. R., & Lewandowski, H. J. (2019). Using think-aloud interviews to characterize model-based reasoning in electronics for a laboratory course assessment. *Physical Review Physics Education Research, 15*, 010140.

Robbins, N., & Heilberger, R. (2011). Plotting Likert and other rating scales. (Paper presented at the Proceedings of the 2011 Joint Statistical Meeting, Miami Beach, FL, USA)

Rosenthal, S. (2018). Motivations to seek science videos on YouTube: Freechoice learning in a connected society. *International Journal of Science Education, Part B, 8*(1), 22–39.

Sandmann, A. (2014). Thinking aloud - analysing thinking, learning and problem-solving processes. Methoden in der naturwissenschaftsdidaktischen Forschung, 179–188.

Sasaki, M. (2013). Introspective methods. A. Kunnan (Ed.), Companion to language assessment. Wiley.

Schwartz, B. (2016). *The paradox of choice: Why more is less: How the culture of abundance robs us of satisfaction*. New York: Ecco Press.

Shruti Goyal, K.P., & Miyapuram, U.L. (2015). Predicting consumer's behavior using eye tracking data. (Paper presented at the Second International Conference on Soft Computing and Machine Intelligence, Hong Kong)

Smith, A. D., Mestre, J. P., & Ross, B. H. (2010). Eye-gaze patterns as students study worked-out examples in mechanics. *Physical Review Special Topics-Physics Education Research, 6*(2), 020118.

Stender, A. (2014). Lesson planning: From knowledge to action: Theoretical development and empirical testing of the transformation model of lesson planning (Vol. 170). Logos Verlag Berlin GmbH.

Stockwell, B. R., Stockwell, M. S., Cennamo, M., & Jiang, E. (2015). Blended learning improves science education. *Cell, 162*(5), 933–936.

Tadbier, A. W., & Shoufan, A. (2021). Ranking educational channels on YouTube: Aspects and issues. *Education and Information Technologies, 26*, 3077–3096.

Tan, E., & Pearce, N. (2011). Open education videos in the classroom: Exploring the opportunities and barriers to the use of YouTube in teaching introductory sociology. *Research in learning technology,* 19.

Tan, E. (2013). Informal learning on YouTube: Exploring digital literacy in independent online learning. *Learning, media and technology, 38*(4), 463–477.

Toffler, A. (1984). *Future shock*. New York: Bantam.

van der Laan, L. N., Hooge, I. T. C., de Ridder, D. T. D., Viergever, M. A., & Smeets, P. A. M. (2015). Do you like what you see? The role of first fixation and total fixation duration in consumer choice. *Food Quality and Preference, 39*, 46–55.

Veith, J., Bitzenbauer, P., & Girnat, B. (2022). Assessing learners' conceptual understanding of introductory group theory using the CI^{2}GT: Development and analysis of a concept inventory. *Education Science*, *12*(376).

Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems, 62*, 1–10.

Welbourne, D. J., & Grant, W. J. (2016). Science communication on YouTube: Factors that affect channel and video popularity. *Public understanding of science, 25*(6), 706–718.

Winkler, B., Bitzenbauer, P.,& Meyn, J.-P. (2021). Quantum physics $\neq$ quantum physics. a survey of researchers' associations. *Physics Education*, *56*(6), 065031.

Winkler, B., Veith, J., & Bitzenbauer, P. (2023). Classical mechanics key topics in physics teacher education: Results of an exploratory mind map study. *European Journal of Educational Research*, *12*(3), 1247–1255.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist, 43*(1), 49–64.

Wolf, K.D., & Kratzer, V. (2015). Explanatory structures in self-created explanatory videos by children. Jahrbuch Medienpädagogik 12: Kinder und Kindheit in der digitalen Kultur, 29–44.

Wu, C.-J., & Liu, C.-Y. (2021). Eye-movement study of high-and lowprior- knowledge students' scientific argumentations with multiple representations. *Physical Review Physics Education Research, 17*(1), 010125.

Authors and Affiliations

Philipp Bitzenbauer[1] · **Tom Teußner[2]** · **Joaquin M. Veith[3]** ·
Christoph Kulgemeyer[4]

[1]   Friedrich-Alexander-Universität Erlangen-Nürnberg, Professur für Didaktik der Physik, Staudtstr. 7, 91058 Erlangen, Germany

[2]   Universität Leipzig, Institut für Didaktik der Physik, Prager Str. 36, 04317 Leipzig, Germany

[3]   Stiftungsuniversität Hildesheim, Institut für Mathematik und Angewandte Informatik, Samelsonplatz 1, 31141 Hildesheim, Germany

[4]   Universität Bremen, Institut für Didaktik der Naturwissenschaften, Otto-Hahn-Allee 1, 28359 Bremen, Germany