Do various groups involved in physics education appreciate the same aspects of physics demonstrations? – A methodological approach

Alexandr Nikitin[1,*], Marie Snětinová[1]

[1] Faculty of Mathematics and Physics, Charles University, V Holešovičkách 2, 180 00 Prague, Czech Republic; alexandr.nikitin@matfyz.cz

This paper presents a novel methodological approach to examining how different groups involved in upper secondary physics education perceive lecture demonstrations. The research utilizes a video-based, mixed-methods design that integrates high-inference rating scales and open-ended qualitative questions. This captures both holistic and analytical evaluations of demonstration quality. The paper focuses on the psychometric properties of the quantitative instrument and the alignment of ratings across four groups: secondary school students, in-service teachers, pre-service physics teachers, and teacher trainers. Initial findings suggest minimal statistically significant differences in how these groups evaluate physics demonstrations, indicating a potential universality in their perceptions. The methodological approach described in this paper offers a framework for assessing experiment focused lecture demonstrations. By providing insights into perceptions of teaching practices during demonstrations, this paper contributes to improving the design and delivery of physics demonstrations that engage diverse audiences and foster conceptual understanding.

Key words:
physics demonstrations, mixed methods design, upper-secondary education, video-study, comparative analysis.

Received 12/2024
Revised 8/2025
Accepted 8/2025

1 Introduction

Experiments are fundamental to both physics as a scientific discipline and physics education (Owen et al., toward more student-centred learning approaches, physics demonstrations continue to play a crucial role in educational settings. Previous research highlights both the benefits and limitations of demonstrations in fostering students' interest and understanding. For instance, some studies report that science demonstrations can raise students' interest in a given field (Lin et al., 2013) and increase students' understanding (Basheer et al., 2017), while others suggest that lecture demonstrations have a negligible impact on students' learning outcomes (Crouch et al., 2004). Despite these mixed findings, demonstrations remain amongst the most common forms of experimentation in many physics classrooms (Seidel et al., 2006), making them a valuable subject of research attention.

Our Department of Physics Education has been conducting physics lecture demonstrations (DEMOs) for upper secondary school students (ISCED 3) for over 30 years. Each year, the DEMOs engage approximately 5 000 students. In 2017/2018, a questionnaire-based study using the Intrinsic Motivation Inventory (Ryan, 1982) examined students' perceptions of these demonstrations (Kácovský & Snětinová, 2021). The results revealed significant variation in students' engagement with different physics topics, which led to a follow-up video-study aimed at identifying specific performance aspects that influence these perceptions.

To address these variations, we previously developed a categorical system[1] for analysing lecturer behaviour during demonstrations, focusing on audience interaction and the use of audiovisual technology (Nikitin, 2021). The results revealed that the various topics performed by different lecturers are very diverse regarding the mentioned aspects (Nikitin et al., 2022). However, the comparison with the data from the Intrinsic Motivation Inventory questionnaire did not provide a satisfactory explanation of the variations in students' perception. This quantitative analysis showed some similarities in the interaction patterns during topics that received lower ratings. Ill-perceived performances were generally less interactive, with the interaction being less varied than in the well-perceived ones. Yet some of the DEMOs did not follow this tendency and even contradicted it. This indicates that there are important aspects of these performances that were not sufficiently covered by the conducted quantitative analysis.

In response, this study adopts a new methodological approach to broaden and deepen our analysis of physics demonstrations. We incorporate an expert evaluation method using high-inference rating scales and engage diverse groups (referred to as *raters* in this paper) related to physics education – including secondary school teachers, students, pre-service teachers, and teacher trainers – to assess demonstration quality. This paper presents a detailed description of the newly developed methodology, offering a replicable framework for future research in science education. Identifying aspects of demonstrations that are valued by different stakeholder groups can help pinpoint the various parameters of the performances that

[1] Applicable to all sorts of lecturer-centred science shows.

influence students' perception. These insights may support the development of more effective teaching practices in physics education.

The paper is structured as follows. Section 2 reviews prior research on physics lecture demonstrations and introduces the specific demonstrations (DEMOs) studied in this work. Section 3 outlines the overall research design, with a detailed description of the video-based evaluation tool used to capture perceptions of demonstration quality. Section 4 focuses on the methodological approach, including the validation of the tool's psychometric properties and the statistical techniques employed to process the quantitative data. Section 5 presents illustrative results from the comparison of rater groups, highlighting the practical application of the methodology. Finally, Section 6 offers conclusions and discusses the implications of this study for future research and teaching practices in physics education.

The primary focus of this paper is the detailed description and validation of the video-based methodology used to evaluate perceptions of physics demonstrations. Comparative findings from the different rater groups are included to demonstrate the practical application of the method and to illustrate its potential for providing meaningful insights into demonstration quality.

2 Research context

2.1 Lecture demonstrations

Lecture demonstrations are widely used in science education as an engaging instructional tool. Their effectiveness in promoting student learning and conceptual understanding has therefore been extensively researched and debated.

Ample evidence across multiple studies supports the use of lecture demonstrations in facilitating cognitive and affective learning outcomes. Austin and Sullivan (2019), Basheer et al. (2017), and Breckler et al. (2013) all reported that demonstrations led to significant improvements in conceptual understanding compared to the initial understanding, academic achievement, and retention of concepts. Crouch et al. (2004) argued that passive observation of demonstrations has a negligible effect, whereas involving student predictions and interaction enhances learning outcomes. This finding is further supported by other research (Manivannan & Meltzer, 2001; Milner-Bolotin et al., 2007; Thornton & Sokoloff, 1998). Furthermore, the control group (without demonstrations) in the study by Breckler et al. (2013) caught up to the experimental group (with demonstrations) after a month, suggesting only a temporary advantage gained from watching demonstrations.

Di Stefano (1996) and Milne and Otieno (2007) highlighted the positive impact of demonstrations on student engagement, interest, and emotional energy, fostering a conducive learning environment. According to Kácovský and Snětinová (2021), these affective gains were particularly evident among students intending to study physics or who felt competent in the subject.

While demonstrations offer potential benefits, their effectiveness depends on various factors identified across multiple studies. Chang and Shieh (2018), Miller et al. (2013), Neo and Yap (2009), and Roth et al. (1997) emphasized the importance of students' prior knowledge, opportunities for prediction and discussion, and explicit instructional guidance in facilitating meaningful learning from demonstrations.

Various studies proposed structured frameworks and strategies to optimize demonstration implementation. Examples include interactive lecture demonstrations (Sokoloff & Thornton, 1997; Zimrot & Ashkenazi, 2007), the survey-question-experiment-recite-reflect-review approach (Chamely-Wiik et al., 2014), using demonstrations as contextual roadmaps (Buncick et al., 2001), ensuring accurate student observation (Miller, 2013), and limiting demonstration duration for better focus (Walton, 2002).

Despite the positive outcomes reported, some studies revealed contrasting findings or limitations. Odom and Bell (2015) found a negative association between teacher demonstrations and student achievement, suggesting demonstrations alone may be insufficient for developing scientific understanding. Thijs and Bosch (1995) observed no significant differences in learning outcomes between teacher demonstrations and small-group practical work. However, in small-group practicals, girls showed a tendency to underperform compared to boys. Rose (2018) noted challenges in delivering effective live demonstrations and suggested that supplementing them with videos accessible outside class time may be more efficient.

Additionally, Sharma et al. (2010) reported learning gains from *interactive lecture demonstrations* that were lower than previously claimed, although still substantial. This finding emphasizes the need for realistic expectations and further research. Moll and Milner-Bolotin (2009) suggest that *interactive lecture demonstrations* have the potential to improve academic achievement. However, their effectiveness depends on implementation strategies, feedback mechanisms, and alignment with assessment practices.

The studies collectively highlight the potential benefits of demonstrations in enhancing student engagement, visualization, and understanding. At the same time, they also emphasize the importance of instructional design, implementation strategies, and consideration of student characteristics and prior

knowledge. Effective lecture demonstrations should be interactive, encourage accurate observation, and be combined with other active learning strategies. Future research should continue to explore ways to optimize the use of demonstrations and investigate their long-term impact on student learning.

2.2 Physics demonstrations for upper secondary school students

The Department of Physics Education at MFF, Charles University, has a long-standing tradition of organizing *Physics demonstrations for upper secondary school students* (DEMOs) (Kácovský & Snětinová, 2021). These lecture demonstrations consist of a thoughtfully selected series of physics experiments accompanied by theoretical explanations. At the time of this research, DEMOs offered seven specialized monothematic performances: Acoustics, Electricity and magnetism, Ionising radiation, Mechanics, Optics, Thermodynamics and molecular physics, and Electromagnetic waves.

DEMOs are held in a university lecture hall and are led by one or two lecturers. Each session lasts 75 minutes and typically draws 60–90 students from multiple schools. Student participation in the experiments is generally limited, with only a few volunteers directly involved. However, focus on fostering student understanding is strong, as lecturers provide both basic and advanced explanations.

One video recording of each performance within DEMOs was obtained during the 2017/2018 school year. To avoid disrupting the usual flow of the sessions, the recordings were taken from a video room located behind the lecture hall. The camera focused on the lecturers' performance, experimentation and presentation. After each session, the lecturers were asked whether the performance had proceeded normally or if it should be recorded again. This option was used only once, when the participants arrived late and the session would otherwise have been significantly shortened.

2.3 Video as a common tool for educational inquiry

Video-based methodologies are widely used in educational research to analyse teaching practices. These studies typically involve systematic coding of classroom interactions and teacher behaviours within recorded lessons. Coders often assign codes from predefined frameworks (e.g., Roth et al., 2006) or rate behaviours using structured scales, such as the Likert scale (e.g., Dalehefte et al., 2009), to evaluate the intensity of specific actions. An alternative approach involves scales with explicit category definitions for behaviour evaluation, such as the assessment of teachers' curriculum knowledge (Wang et al., 2023). Coding systems vary in their level of interpretation, ranging from high-inference methods, which require subjective judgment, to low-inference methods that rely on more objective classifications (Dalehefte et al., 2009). These methods facilitate the analysis of both the observable and the contextual aspects of teaching. Some studies even examinee multiple dimensions of behaviour within the same video segment.

Video-based methodologies have proven effective across various research designs. Seidel and Prenzel (2006) used time sampling to analyse physics lessons and demonstrated stable teaching patterns through a coding framework with high inter-coder reliability. Jewitt (2012) emphasized the capacity of video to capture multimodal classroom interactions, allowing for detailed analysis of gestures, expressions, and speech. However, he also noted challenges such as camera effects and ethical concerns. Zhang et al. (2011) highlighted the benefits of self-produced videos in teacher reflection, contrasting them with published videos that model best practices, although technical limitations remained an obstacle. Similarly, Vondrová and Žalská (2018) found that while pre-service teachers could identify mathematical phenomena in videos, their interpretive skills showed limited progress.

Studies also support the role of video in teacher professional development. Simpson et al. (2018) found that guided observation programs improved teachers' focus on student engagement and reasoning. However, interpretation skills still required further training. Lebak (2023) demonstrated that video-based pedagogical action research fosters systematic reflection, helping teachers address instructional challenges. Together, these findings underscore the potential of video for enhancing teacher awareness and instructional improvement, provided that structured guidance and sustained practice are included. In recognition of this impact, we have incorporated a dedicated video study into our research design.

The methodologies outlined above align closely with the aims of this research, particularly the use of structured scales and coding frameworks to analyse teacher behaviours. Building on these approaches, this study adopts a multi-dimensional analysis that leverages video data to capture both the observable and contextual aspects of teaching. By integrating video reflection into a broader research design, this study aims to address existing gaps in interpretive training and contribute to the refinement of video-based methodologies.

3 Overall research design

From a methodological perspective, the research design can be described as an observational video-study using high-inference rating scales for expert evaluation. The observation in our research is open, non-participatory, partially structured and conducted in a natural setting (although captured on video).

To comprehensively address the aspects that may influence students' perceptions, we evaluated the recordings from two perspectives – as a *whole topic* and as many monothematic *short sections* into which these performances are divided.

The research includes four groups of raters (21 raters each) related to upper-secondary school physics education – upper-secondary school students, upper-secondary school teachers, pre-service physics teachers and physics teacher trainers.

3.1 Assessing whole topic and short sections

One video recording of each topic, approximately 75 mins long, was analysed. These recordings are referred to as *whole topics*, and one unique set of rating scales was developed for their assessment. Each *whole topic* was evaluated by twelve distinct raters.

Furthermore, the recording of each whole topic was divided into several monothematic short videos, typically about fourteen per topic. Here, monothematic refers to a short video centred on a single topic or physics phenomenon, with an average length of about six minutes. These are referred to as *short sections*, and a specific set of scales was developed to assess them.

Of the short sections, 55 were classified as *mainly experimental*, containing only minimal theoretical explanation. An additional 29 sections were identified as *mixed*, combining experiments with theory. Finally, 16 sections were designated as *mainly theoretical*, involving only minimal experimental activity. Each of these sections was assessed by at least nine distinct raters.

3.2 Rating scales

Two sets of rating scales were developed, each tailored to the nature of the performances being studied and to the specifics of evaluating either a *short section* (one set of scales), or a *whole topic* (the other set of scales). The following sections list the scales for both types of recordings, beginning with short sections.

Short section scales:

Atmosphere in the auditorium: Assesses how well the lecturer maintains a focused and engaging atmosphere in the audience, balancing attention, eye contact, humour, and age appropriateness.

Experiment clarity: Evaluates how clearly the lecturer conveys the purpose and results of experiments, making them accessible and logically connected to theoretical points.

Visibility: Measures the visibility of demonstrations and equipment for all audience members, including diagrams or camera use if needed.

Speech clarity: Assesses how understandable the lecturer's explanations are, ensuring that scientific terms and complex ideas are conveyed clearly to a lay audience.

Overall impression of the lecturer's performance: Summarizes the lecturer's effectiveness, confidence, and ability to engage the audience through the presentation.

The designed scales consist of five points, with detailed descriptions provided for the 1st, 3rd, and 5th point. The *Experiment clarity* and *Visibility* scales also contain the point **N**, denoting "not relevant" or "not judgeable" (e.g., evaluating visibility of an experiment in a mainly theoretical section). *Visibility* scale is provided here as an example:

Scale: Visibility

5 Demonstrated or measured phenomena are clearly visible, provably observable. The equipment is sufficiently large and visible to the audience, even from the back rows, or a camera is appropriately used. Diagrams and drawings are sufficiently visible.

3 Some deficiencies reduce the visibility of the experiment – the effect is less noticeable or not equally visible to all viewers.

1 The experiment is poorly visible to some viewers; the equipment is too small.

N Not relevant for this section.

Each scale is accompanied by a written commentary related to the aspects measured. This commentary is mandatory if the rater selects a non-extreme point of the scale (points **2**, **3**, **4** or **N**), in which case the rater is required to justify their choice.

Whole show scales:

Introduction and establishing contact: Evaluates the lecturer's ability to quickly establish rapport with the audience and create a relaxed, interactive environment in the first five minutes.

Interaction with the audience: Evaluates the lecturer's efforts to engage the audience consistently through questions, discussions, and responses to inquiries.

Atmosphere in the auditorium: Measures the overall engagement and atmosphere created by the lecturer during the show.

Utilization of motivation: Assesses the lecturer's use of motivational techniques (like surprise or problem questions) to maintain interest and emphasize relevance of the content.

Overall logical structure: Evaluates the logical coherence and flow of the presentation, particularly the connection between theory and experiments.

***Overall subjective impression*:** Summarizes personal responses to the lecturer's style, engagement level, and perceived value of the performance, using both a rating scale and open-ended questions.

The whole show scales are designed in the same way as the short section scales (see the example of *Visibility* above), except for *Overall subjective impression*.

Unlike the other scales, the *Overall subjective impression* consists of a series of scales and questions intended for two purposes: (a) to estimate the rater's impression of the performance with an emphasis on their subjective perception, and (b) to gather information about the rater's attitude towards experiments and their evaluation strictness. The second goal is achieved through ten statements rated on a four-point Likert-type scales (with the neutral point being purposely omitted). For example:

- I felt immersed in the plot of the performance.
- The lecturer's style of presentation is very interesting to me.
- I enjoyed the performance.

These scales are followed by four open-ended questions:

- The moment that interested me the most during the DEMOs (e.g. specific experiments, presentation methods, explanations, lecturer's reactions, . . .)
- I think the performance could be improved by. . .
- What I appreciate the most about the presenter. . .
- Is there anything you would like to add to the performance that you could not express in the previous items?

The *Overall subjective impression* concludes the observation sheet and serves as a final reflection point for the raters.

3.3 Raters, rater groups and allotment of videos to raters

Collectively, 84 raters, evenly distributed amongst four different groups of respondents involved in Czech physics instruction, participated in this research. These groups are *in-service physics teachers* (I), *upper-secondary school students* (S), *pre-service physics teachers* (P), and *physics teacher trainers* (T). One of the objectives of this study is to ascertain whether these groups involved in physics education value the same aspects of lecture demonstrations.

Each rater evaluated eleven *short sections* and one *whole topic* using the previously described scales. The short sections were systematically allotted to the raters according to the following rules:

- evaluation of the *whole topic* excluded evaluation of *sections* from that particular performance;
- the distribution of *sections* assigned to each rater reflected the distribution of all sections (6/11 *mainly experimental*, 3/11 *mixed*, and 2/11 *mainly theoretical*);
- thematic diversity was ensured (*sections* originated from several topics of DEMOs);
- there was a sufficient overlap both among raters and among sections.

Raters underwent a training session (with a recording available afterwards), and received codebooks describing the fundamental principles of video evaluation and detailed explanations of the scales used.

Data from *in-service physics teachers* and *upper-secondary school students* was gathered during summer and autumn of 2022, while data from *pre-service physics teachers* and *physics teacher trainers* was gathered during summer and autumn of 2023.

4 Methodological approach

Due to the complex nature of the designed research tool, several methodological considerations need to be addressed. Firstly, we verify that the attitude statements display reasonable psychometric properties, as they are used to improve the objectivity of the respondent's answers. We also examine psychometric properties of the individual responses to both the short section and whole show scales. Finally, we describe the procedure for adjusting rater's responses according to their attitude towards experimenting and present its results.

4.1 Psychometric analysis of attitude statements

As previously mentioned, the *overall subjective impression of the show* covers ten items with 4-point Likert scales (*disagree – rather disagree – rather agree – agree*). These items aim to estimate respondents' strictness and their attitude toward experimenting, as both are important factors that may influence their ratings. Table 1 presents the complete list of these items.

Table 1: Attitude statements used to estimate respondent's attitude towards experimenting

item code	reverse worded	item wording
p1	yes	The lecturer's style of demonstrating experiments and explanations does not sit well with me.
p2		I was intrigued by the performance the whole time.
p3		I consider the time spent watching and evaluating the performance to be meaningfully spent.
p4	yes	I was bored while watching the performance.
p5		I felt immersed in the show.
p6		The lecturer's style of presentation is very interesting to me.
p7		I enjoyed the show.
p8	yes	At times, I lost my attention during the performance.
p9	yes	The performance was not interesting to me.
p10	yes	I did not find watching and judging the performance useful.

Responses to the reverse worded items were recoded so that more positive responses were assigned higher values scaling from 1 to 4. Therefore, the increasing value reflects both a respondent's better attitude towards experimentation and a respondent who is more benevolent in their ratings. We refer to this feature as *respondent's bias* (or simply *bias*).

Inspecting parameters of the distribution of responses to these items in Table 2 reveals that the answers are negatively skewed, with responses concentrating around the more positive values of the scales.

Figure 1 shows a heatmap of Pearson's correlations between these items. Items p3, p9 and p10 show weaker correlations with the remaining items. Analysis of reliability (Table 2) confirms this pattern: dropping item p10 would increase the internal consistency of the scale. Subsequently, removing item p10 identifies item p9 for removal and removing item p9 in turn identifies item p3. Eliminating these three items results in a scale exhibiting internal consistency of 0.934 in Cronbach's alpha.

Any further analysis of the attitude statements used to estimate *respondent's bias* is based only on items p1, p2 and p4–p8.

Table 2: Analysis of reliability of attitude statements (marked by their item code p1–p10)

item	*mean*	*SD*	*min*	*Q1*	*Q2*	*Q3*	item-rest correlation	Cronbach's α
scale	**3.30**	**0.652**					–	**0.914**
								If item dropped
p1	3.58	0.746	1	3	4	4	0.747	0.903
p2	3.08	0.903	1	3	3	4	0.832	0.896
p3	3.48	0.683	1	3	4	4	0.537	0.913
p4	3.33	0.892	1	3	4	4	0.789	0.899
p5	3.06	0.943	1	2	3	4	0.809	0.897
p6	3.25	0.898	1	3	3	4	0.714	0.903
p7	3.35	0.841	1	3	4	4	0.864	0.895
p8	2.79	1.114	1	2	3	4	0.694	0.907
p9	3.51	0.826	1	3	4	4	0.557	0.912
p10	3.58	0.762	1	3	4	4	0.319	0.924

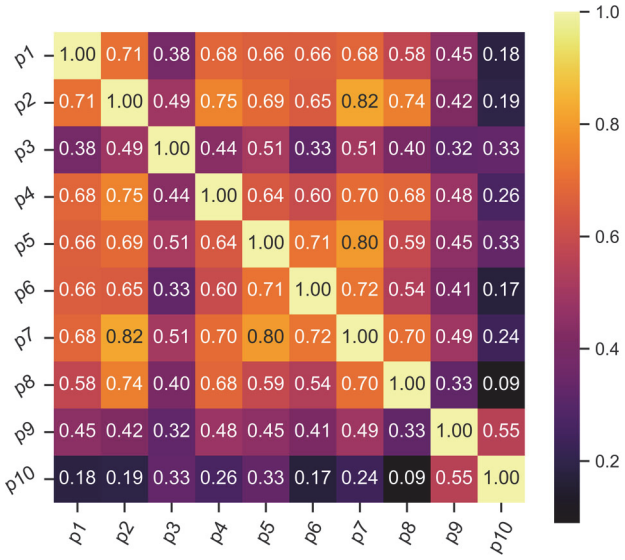

Fig. 1: Correlation heatmap of the attitude statements (marked by their item code p1–p10)

4.2 Descriptives of the short sections and whole shows scales

All the scales used for both – short sections and whole shows – consist of 5 points with higher values corresponding to more positive ratings. The mean and median (Q2) for the whole show scales in Table 3 suggest that the responses are generally positive. Because the median is already equal to the highest rating on the scales, the 3rd quartile (Q3) and maximum are equal to it and therefore omitted from the table.

Table 3: Whole show scale descriptives and reliability analysis

$N = 85$	*mean*	*SD*	*min*	*Q1*	*Q2*	item-rest correlation	Cronbach's α
scale	**4.45**	**0.666**				–	**0.845**
							If item dropped
introduction	4.40	1.01	1	4	5	0.678	0.811
interaction	4.33	0.97	1	4	5	0.804	0.767
atmosphere	4.39	0.90	2	4	5	0.748	0.785
motivation	4.55	0.72	3	4	5	0.756	0.792
logical structure	4.71	0.55	3	5	5	0.317	0.881

The fact that more than 50% of responses correspond to the highest rating provides evidence of the overall quality of DEMOs. However, it also provides a methodological challenge, as the responses do not allow for clear differentiation among the shows with good ratings.

Regarding reliability analysis of these scales, the *overall logical structure of the show* correlates the least with the remaining scales, and its removal increases Cronbach's α to 0.88. This result is not surprising, as this scale measures qualitatively different and more objective aspect than the others, which mainly capture subjective experiences or perceptions, which can vary widely across individuals based on personal preferences and emotional responses.

The high internal consistency suggests that a combination of the remaining four scales can reasonably be used as a measure of the show's quality. Further inspection of the heatmap in Figure 2 shows that these scales also correlate well with the attitude scales (apart from *overall logical structure of the performance*).

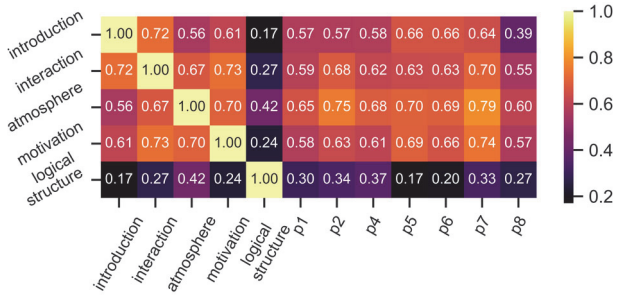

Fig. 2: Correlation heatmap of the whole show scales with the attitude statements p1–p8

Table 4: Short section scales descriptives

	N	*mean*	*SD*	*min*	*Q1*	*Q2*
atmosphere	937	4.38	0.92	1	4	5
experiment clarity	787	4.56	0.86	1	4	5
visibility	806	4.34	1.01	1	4	5
speech clarity	937	4.64	0.73	1	5	5
overall impression	937	4.41	0.85	1	4	5

As shown in Table 4, the scales for short sections are skewed towards the positive scale ratings as well. More than 50% of responses represent the highest values of the scales. Since the scales *visibility* and *experiment clarity* allow for the choice *N* (non-relevant), these are treated as "missings" in the quantitative analysis, which explains the lower number of responses.

The correlation matrix in Figure 3 shows that the *short section scales* correlate significantly less at the level of individual responses than the *whole show scales*. This outcome is expected, as the short sections are significantly more varied regarding quality of the assessed aspects. Since the short section scales are never analysed as a single aggregated scale, this does not present a problem.

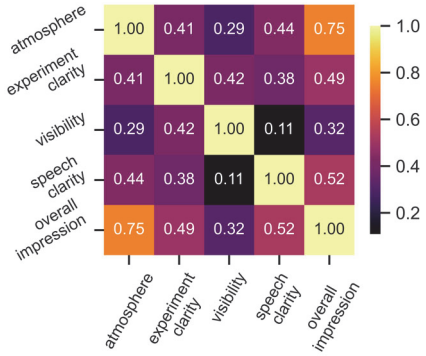

Fig. 3: Short section scales correlation heatmap

4.3 Response level correction

We assume that each respondent strives to be as objective as possible. However, the way they use the scales differs based on their individual *bias*. A person that has better attitude towards experimenting in physics (higher *bias*) tend to be more benevolent in their ratings. Therefore, a correction of the numerical

scale levels is performed separately for each rater and applied to all their responses. This correction was carried out as follows:

1) Each respondent was assigned an average score of their responses to the attitude statements – an estimate of their *bias*.
2) Each respondent was also assigned an average score of their rating of the *whole show*.[2]
3) For each whole show, a linear regression of the respondent's average rating on their bias was estimated using weighted ordinary least squares with 1 000 bootstraps.
4) Each respondent was then assigned a residual from this model. The reasoning is that respondents above the regression line are more positive in their responses than they should be (according to their bias), and thus use the scales more benevolently, while those below the regression line are more critical. The histogram of these residuals shown in Figure 4 reveals that these corrections are relatively small for majority of respondents.
5) Each response for each respondent is lowered by this value (responses of those using the scales more benevolently are lowered, and vice versa).
6) A scale-index estimate is computed as the mean of these responses with corrected levels, the standard error of the mean is used as estimation error.

This approach considers both the multilevel research design and respondent *bias*.

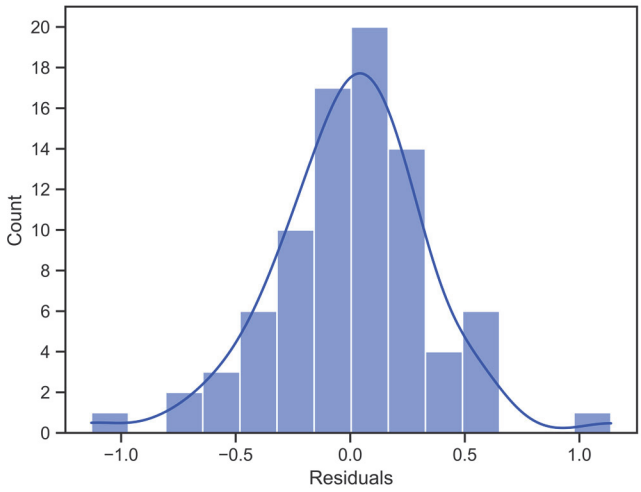

Fig. 4: Histogram of residuals used for response-level correction (with a kernel density estimation)

In simpler terms, if a respondent indicates a strong bias but their responses to *whole show scales* do not reflect it (e.g. they are too benevolent given their bias), their responses are adjusted according to the "general consensus about the whole show" as represented by the linear model. Thus, the correction does not aim to eliminate *respondent's bias* entirely, but rather to estimate how each respondent uses the designed rating scales.

5 Illustrative results – comparison of rater groups

The theory of incomplete block designs (Dey, 2010) is used to compare various rater groups.

In essence, the responses to the individual *short section scales* with performed *response level correction* are analysed using a mixed-effects linear model, with individual short sections as fixed factors[3] and individual respondents as random factors.[4] Estimation of the random factors of this model provides an estimate of each respondents' effect on the ratings of the particular short section scale.

According to Vonesh et al. (1996), R^2 for mixed-effects models can be categorised into two types: marginal R^2 and conditional R^2. Marginal R^2 relates to the variance explained by the fixed factors, while conditional R^2 relates to the variance explained by both the fixed and random factors (Nakagawa

[2] Excluding the scale *overall logical structure of the performance* as discussed above.
[3] *Treatments* in the language of block designs.
[4] *Blocks* in the language of block designs.

et al., 2013). Table 5 states these goodness-of-fit statistics for the mixed-effects models for each scale. The results show that including respondents as random effects in the models significantly improves their explanatory capabilities, with the full models explaining about 90% of variance in the responses. This proves, that the response level correction described in Subchapter 4.3 does not remove the influence of individual raters; rather, it serves as a 'calibration' of how the raters interpret the definitions of the rating scales.

Table 5: Goodness-of-fit of the mixed-effects models for individual short section scales

	overall impression	atmosphere	speech clarity	speech clarity	visibility
R^2_{marg}	0.504	0.529	0.423	0.431	0.469
R^2_{cond}	0.923	0.926	0.911	0.898	0.912

Descriptive statistics of the respondent effects estimated from these models are presented in Table 6 for each respondent group separately. The table shows that the means for *in-service teachers* (I), and *physics teacher trainers* (T) are negative, while the means for *upper-secondary students* (S) and *pre-service teachers* (P) are positive. This suggests that I and T tend to be more critical towards the DEMOs than S and P. A brief look at the boxplots of the respondent effects in Figure 5 further shows that the group distributions overlap significantly.

Table 6: Descriptives of the estimated respondent effects on the short scale ratings split by the respondent group (I – in-service physics teachers; P – pre-service physics teachers; S – upper secondary school students; T – physics teacher trainers)

scale	respondent group	mean	SEM	min	*Q1*	*Q2*	*Q3*	max
overall impression	I	−0.03	0.11	−1.50	−0.13	0.02	0.29	0.95
	P	0.11	0.08	−0.67	−0.03	0.14	0.38	0.69
	S	0.06	0.10	−0.68	−0.24	0.06	0.36	0.82
	T	−0.14	0.12	−1.47	−0.41	−0.09	0.08	0.86
atmosphere	I	−0.05	0.11	−1.38	−0.21	0.16	0.34	0.56
	P	0.13	0.08	−0.84	0.00	0.21	0.41	0.57
	S	0.09	0.12	−0.91	−0.21	0.09	0.44	1.13
	T	−0.17	0.14	−1.58	−0.36	−0.06	0.10	1.19
speech clarity	I	−0.11	0.09	−1.14	−0.26	−0.01	0.28	0.43
	P	0.12	0.07	−0.60	−0.06	0.11	0.38	0.70
	S	0.12	0.10	−0.49	−0.30	0.10	0.27	1.15
	T	−0.13	0.09	−1.12	−0.37	−0.07	0.10	0.66
speech clarity	I	−0.10	0.12	−1.71	−0.19	0.00	0.17	0.61
	P	0.12	0.09	−0.95	0.05	0.20	0.33	0.62
	S	0.17	0.11	−0.86	−0.10	0.18	0.62	0.87
	T	−0.20	0.12	−1.32	−0.38	−0.19	−0.01	0.65
visibility	I	−0.04	0.11	−1.25	−0.26	0.03	0.30	0.59
	P	0.21	0.08	−0.74	0.06	0.17	0.40	0.83
	S	0.02	0.10	−1.28	−0.24	0.12	0.33	0.81
	T	−0.18	0.10	−1.28	−0.37	−0.02	0.15	0.47

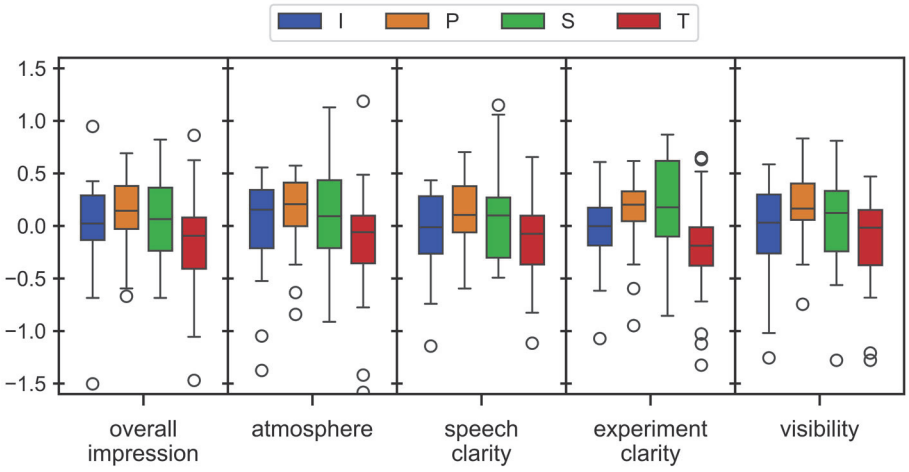

Fig. 5: Boxplots of the estimated respondent effects split by the respondent group (I – in-service physics teachers; P – pre-service physics teachers; S – upper secondary school students; T – physics teacher trainers)

The results of Welch's one-way ANOVAs in Table 7 confirm that the distributions overlap significantly, as only *visibility* shows statistically significant differences ($p < 0.05$) between the respondent groups. According to the Games-Howell post hoc test, the only significant difference in *visibility* is between T and P, with P being more positive in their ratings (mean difference 0.39, Cohen's $d \approx 1.24$).

Table 7: One-way Welch's ANOVA for the effect of respondent groups on individual short section scales

scale	*F*	*df1*	*df2*	*p*
overall impression	1.16	3	44.0	0.335
atmosphere	1.34	3	43.8	0.272
speech clarity	2.47	3	44.1	0.074
experiment clarity	2.60	3	44.1	0.064
visibility	3.19	3	44.1	0.033

There are practically no significant differences between the quantitative responses of different rater groups. Although *physics teacher trainers* appear somewhat more critical, as one might expect, and *pre-service teachers* somewhat more benevolent, these differences are statistically insignificant, with the sole exception of *visibility*.

6 Conclusions

Our research methodology is unique in several key aspects:

1. **Comprehensive performance evaluation:**

   Unlike approaches that focus solely on selected key moments, we assess entire performances by systematically segmenting videos and evaluating each segment without exception. This eliminates the potential bias caused by selectively choosing specific parts for analysis.

2. **Systematic segmentation and consistency:**

   The videos are divided into segments according to a systematic framework, ensuring uniformity and objectivity in segment selection. Every segment is assessed using the same predefined scales, regardless of its content. This standardization allows for direct comparisons between different parts of the video and ensures consistency in data collection.

3. **Detailed and explicit evaluation scales:**

   Our five-point scales include detailed descriptions for three anchor points, expressed in several sentences rather than just a few words. This level of detail reduces ambiguity and makes the scales more explicit compared to less defined metrics or those relying heavily on subjective coder interpretations.

4. **Qualitative insights:**

   Raters are encouraged to supplement numerical scores with written justifications. This allows for a deeper reflection and adds qualitative depth to the data collected, extending beyond purely numerical analysis.

5. **Emphasis on systematic, detailed, and standardized assessment:**

   Our approach prioritizes a methodical and highly detailed evaluation process, creating a standardized framework that promotes more objective and accurate assessments of video performances. It should be noted, however, that this approach may require greater time investment and coordination from raters.

Finally, it is worth noting that our study does not rely on professional perspectives, as even the students participating in the research are not experts. This allows for a broader applicability of our methodology beyond expert-level evaluations, making it accessible and relevant to a wider audience.

This novel methodology distinguishes itself from existing approaches by offering greater accuracy and objectivity in video analysis for educational research, though it requires a higher investment of resources. Our findings indicate that the proposed methodology, together with the evaluation tool for assessing lecture demonstrations, successfully integrates quantitative and qualitative approaches. Initial analysis of the quantitative data confirms the appropriateness of the selected rating scales, demonstrating satisfactory psychometric properties. However, the validity of the qualitative component remains to be examined in further research.

The findings also suggest that various groups of raters (students, physics teachers, pre-service physics teachers, and physics teacher trainers) generally value similar aspects of DEMOs. This points to a degree of consistency in preferences across different roles in physics education. While teacher trainers tend to be more critical in their assessments than students, the differences between the groups are statistically insignificant in most cases. The only significant differences appeared in the visibility of DEMOs, where pre-service teachers were more positive in their ratings compared to teacher trainers.

The minimal differences observed across rater groups likely reflect a shared understanding of what makes lecture demonstrations effective – qualities such as clarity, visibility, and conceptual relevance are widely appreciated in physics education, regardless of role or experience level. This alignment may result from common educational experiences and professional norms that emphasize these features. While teacher trainers tend to be slightly more critical, likely due to their greater pedagogical expertise, the overall consistency indicates that effective demonstrations share broadly recognized characteristics.

To interpret this cautiously, analysing raters' open-ended comments (not covered in this paper) could help explain why more experienced individuals in physics education tend to be stricter. For instance, teacher trainers may detect physics inaccuracies that upper secondary students miss, or they may focus more strongly on evaluating the lecturer's pedagogical content knowledge.

Looking ahead, we aim to investigate how these aspects influence students' perceptions of physics demonstrations by linking the collected data with previously published results from the Intrinsic Motivation Inventory questionnaire (Kácovský & Snětinová, 2021). This approach will help identify which specific parameters of demonstrations contribute most to positive student evaluations, thereby enabling us to optimize the performance of DEMOs.

Acknowledgment

The present work was supported by the Specific University Research Project SVV 260 828 and by Charles University in Prague, project GA UK No. 62223.

References

Austin, S. R. P., & Sullivan, M. (2019). How are we performing? Evidence for the value of science shows. *International Journal of Science Education, Part B*, *9*(1), 1–12. https://doi.org/10.1080/21548455.2018.1532620

Basheer, A., Hugerat, M., Kortam, N., & Hofstein, A. (2017). The effectiveness of teachers' use of demonstrations for enhancing students' understanding of and attitudes to learning the oxidation-reduction concept. *Eurasia Journal of Mathematics Science and Technology Education*, *13*(3), 555–570. https://doi.org/10.12973/eurasia.2017.00632a

Breckler, J. L., Christensen, T., & Sun, W. (2013). Using a physics experiment in a lecture setting to engage biology students with the concepts of poiseuille's law. *CBE—Life Sciences Education*, *12*(2), 262–273. https://doi.org/10.1187/cbe.12-08-0129

Buncick, M. C., Betts, P. G., & Horgan, D. D. (2001). Using demonstrations as a contextual road map: Enhancing course continuity and promoting active engagement in introductory college physics. *International Journal of Science Education*, *23*(12), 1237–1255. https://doi.org/10.1080/09500690010025030

Chamely-Wiik, D. M., Haky, J. E., Louda, D. W., & Romance, N. (2014). SQER3: An instructional framework for using scientific inquiry to design classroom demonstrations. *Journal of Chemical Education*, *91*(3), 329–335. https://doi.org/10.1021/ed300689n

Chang, W., & Shieh, R. S. (2018). A study of the conceptual comprehension of electric circuits that engineer freshmen display. *European Journal of Physics*, *39*(4), 045705. https://doi.org/10.1088/1361-6404/aab6e1

Crouch, C., Fagen, A. P., Callan, J. P., & Mazur, E. (2004). Classroom demonstrations: Learning tools or entertainment? *American Journal of Physics*, *72*(6), 835–838. https://doi.org/10.1119/1.1707018

Dalehefte, I. M., Rimmele, R., Prenzel, M., Seidel, T., Labudde, P., & Herweg, C. (2009). Observing instruction "next-door": a video study about science teaching and learning in Germany and Switzerland. In T. Janík & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 83–99). Waxmann.

Dey, A. (2010). *Incomplete block designs*. World Scientific.

Di Stefano, R. (1996). Preliminary IUPP results: Student reactions to in-class demonstrations and to the presentation of coherent themes. *American Journal of Physics*, *64*(1), 58–68. https://doi.org/10.1119/1.18293

Jewitt, C. (2012). *An introduction to using video for research (National Centre for Research Methods Working Paper 03/12)*. Institute of Education.

Kácovský, P., & Snětinová, M. (2021). Physics demonstrations: who are the students appreciating them? *International Journal of Science Education*, *43*(4), 529–551. https://doi.org/10.1080/09500693.2020.1871526

Lebak, K. A. (2023). Utilizing video-based pedagogical action research to transform teacher practice in elementary, middle and high school classrooms. *Journal of Inquiry and Action in Education*, *11*(2). Retrieved from https://digitalcommons.buffalostate.edu/jiae/vol11/iss2/2

Lin, H.-s., Hong, Z.-R., & Chen, Y.-C. (2013). Exploring the development of college students' situational interest in learning science. *International Journal of Science Education*, *35*(13), 2152–2173. https://doi.org/10.1080/09500693.2013.818261

Manivannan, K., & Meltzer, D. (2001). *Use of in-class physics demonstrations in highly interactive format.* Paper presented at Physics education research conference 2001, Rochester, New York. https://doi.org/10.1119/perc.2001.pr.011

Miller, K. (2013). Use demonstrations to teach, not just entertain. *The Physics Teacher*, *51*(9), 570–571. https://doi.org/10.1119/1.4830081

Miller, K., Lasry, N., Chu, K., & Mazur, E. (2013). Role of physics lecture demonstrations in conceptual learning. *Physical Review Special Topics – Physics Education Research*, *9*(2), 020113. https://doi.org/10.1103/PhysRevSTPER.9.020113

Milne, C., & Otieno, T. (2007). Understanding engagement: Science demonstrations and emotional energy. *Science Education*, *91*(4), 523–553. https://doi.org/10.1002/sce.20203

Milner-Bolotin, M., Kotlicki, A., & Rieger, G. (2007). Can students learn from lecture demonstrations? Journal of College Science Teaching, *36*(4), 45–49. https://www.jstor.org/stable/42992942

Moll, R. F., & Milner-Bolotin, M. (2009). The effect of interactive lecture experiments on student academic achievement and attitudes towards physics. *Canadian Journal of Physics*, *87*(8), 917–924. https://doi.org/10.1139/P09-048

Nakagawa, S., Schielzeth, H., & O'Hara, R. B. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Neo, C. S., & Yap, K. C. (2009). The effect of classroom demonstrations based on conceptual change instruction on students' understanding of electromagnetism and electromagnetic induction. In M. Kim, S. W. Hwang & A.-L. Tan (Eds.), *Proceedings of the International Science Education Conference 2009* (pp. 1346–1386). Singapore: National Institute of Education.

Nikitin, A. (2021). *Fyzikální pokusy pro střední školy – videostudie* [Masters, Charles University]. http://hdl.handle.net/20.500.11956/127719

Nikitin, A., Kácovský, P., & Snětinová, M. (2022). Looking for parameters of students-lecturer interaction that influence how students perceive physics demonstrations. In J. Ondruška, Ľ. Valovičová & Ľ. Zelenický (Eds.), Didactic Transfer of Physics Knowledge Through Distance Education, *AIP Conference Proceedings*, *2458*(1). AIP Publishing LLC. https://doi.org/10.1063/5.0078259

Odom, A. L., & Bell, C. V. (2015). Associations of middle school student science achievement and attitudes about science with student-reported frequency of teacher lecture demonstrations and student-centered learning. *International Journal of Environmental and Science Education*, *10*(1), 87–97.

Owen, S., Dickson, D., Stanisstreet, M., & Boyes, E. (2008). Teaching physics: Students' attitudes towards different learning activities. *Research in Science & Technological Education*, *26*(2), 113–128. https://doi.org/10.1080/02635140802036734

Rose, T. M. (2018). Lessons learned using a demonstration in a large classroom of pharmacy students. *American Journal of Pharmaceutical Education*, *82*(9), 6413. https://doi.org/10.5688/ajpe6413

Roth, W.-M., McRobbie, C. J., Lucas, K. B., & Boutonné, S. (1997). Why may students fail to learn from demonstrations? A social practice perspective on learning in physics. *Journal of Research in Science Teaching*, *34*(5), 509–533. https://doi.org/10.1002/(SICI)1098-2736(199705)34:5¡509::AID-TEA6¿3.0.CO;2-U

Roth, K. J., Druker, S. L., Garnier, H. E., Lemmens, M., Chen, C., Kawanaka, T., Rasmussen, D., Trubacova, S., Warvi, D., Okamoto, Y., Stigler, J., & Gallimore, R. (2006). *Teaching science in five countries: Results from the TIMSS 1999 video study. Statistical analysis report. NCES 2006-011.* ED Pubs.

Ryan, R. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, *43*(3), 450–461. https://doi.org/10.1037//0022-3514.43.3.450

Seidel, T., & Prenzel, M. (2006). Stability of teaching patterns in physics instruction: Findings from a video study. *Learning and Instruction*, *16*(3), 228–240. https://doi.org/10.1016/j.learninstruc.2006.03.002

Seidel, T., Prenzel, M., Rimmele, R., Dalehefte, I. M., Herweg, C., Kobarg, M., & Schwindt, K. (2006). Blicke auf den physikunterricht. Ergebnisse der IPN videostudie. *Zeitschrift für Pädagogik*, *52*(6), 799–821.

Sharma, M., Johnston, I., Helen, J., Varvell, K., Gordon, R., Andrew, H., Stewart, C., Ian, C., & Thornton, R. (2010). Use of interactive lecture demonstrations: A ten year study. *Physical Review Special Topics. Physics Education Research*, *6*, 020119. https://doi.org/10.1103/PhysRevSTPER.6.020119

Simpson, A., Vondrová, N., & Žalská, J. (2018). Sources of shifts in pre-service teachers' patterns of attention: The roles of teaching experience and of observational experience. *Journal of Mathematics Teacher Education*, *21*(6), 607–630. https://doi.org/10.1007/s10857-017-9370-6

Sokoloff, D. R., & Thornton, R. K. (1997). Using interactive lecture demonstrations to create an active learning environment. *The Physics Teacher*, *35*(6), 340–347. https://doi.org/10.1119/1.2344715

Thijs, G. D., & Bosch, G. M. (1995). Cognitive effects of science experiments focusing on students' preconceptions of force: a comparison of demonstrations and small-group practicals. *International Journal of Science Education*, *17*(3), 311–323. https://doi.org/10.1080/0950069950170304

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, *66*(4), 338–352. https://doi.org/10.1119/1.18863

Vondrová, N., & Žalská, J. (2018). Ability to notice mathematics specific phenomena: What exactly do student teachers attend to? *Orbis scholae*, *9*(2), 77–101. https://doi.org/10.14712/23363177.2015.81

Vonesh, E. F., Chinchilli, V. M., & Pu, K. (1996). Goodness-of-Fit in generalized nonlinear mixed-effects models. *Biometrics*, *52*(2), 572–587. https://doi.org/10.2307/2532896

Walton, P. H. (2002). On the use of chemical demonstrations in lectures. *University Chemistry Education*, *6*(1), 22–27.

Wang, J., Wang, Y., Wipfli, K., Thacker, B., & Hart, S. (2023). Investigating learning assistants' use of questioning in online courses about introductory physics. *Physical Review Physics Education Research*, *19*(1), 010113. https://doi.org/10.1103/PhysRevPhysEducRes.19.010113

Zhang, M., Lundeberg, M., Koehler, M. J., & Eberhardt, J. (2011). Understanding affordances and challenges of three types of video for teacher professional development. *Teaching and Teacher Education*, *27*(2), 454–462. https://doi.org/10.1016/j.tate.2010.09.015

Zimrot, R., & Ashkenazi, G. (2007). Interactive Lecture Demonstrations: a tool for exploring and enhancing conceptual change. *Chemistry Education Research and Practice*, *8*(2), 197–211. https://doi.org/10.1039/B6RP90030E