frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 19 October 2023
DOI 10.3389/feduc.2023.1162499

Check for updates

OPEN ACCESS

EDITED BY
Hendrik Lohse-Bossenz,
University of Education Heidelberg, Germany

REVIEWED BY
Ferman Konukman,
Qatar University, Qatar
Matthias Baumgartner,
University of Teacher Education St. Gallen, Switzerland

*CORRESPONDENCE
Christian Leukel
✉ christian.leukel@ph-freiburg.de

RECEIVED 09 February 2023
ACCEPTED 15 September 2023
PUBLISHED 19 October 2023

CITATION
Leukel C, Leuders T, Bessi F and Loibl K (2023) Unveiling cognitive aspects and accuracy of diagnostic judgments in physical education teachers assessing different tasks in gymnastics.
Front. Educ. 8:1162499.
doi: 10.3389/feduc.2023.1162499

COPYRIGHT
© 2023 Leukel, Leuders, Bessi and Loibl. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Unveiling cognitive aspects and accuracy of diagnostic judgments in physical education teachers assessing different tasks in gymnastics

Christian Leukel[1,2]*, Timo Leuders[1], Flavio Bessi[3] and Katharina Loibl[1]

[1]University of Education Freiburg, Freiburg, Germany, [2]Bernstein Center Freiburg, University of Freiburg, Freiburg, Germany, [3]Department of Sport Science, University of Freiburg, Freiburg, Germany

Introduction: Diagnostics is an essential part of teachers' profession. We investigated judgment accuracy and cognitive processes underlying judgment formation in physical education teachers who observed tasks in gymnastics, and compared teachers with gymnastics trainers as a reference group.

Methods: Teachers and trainers judged performance of prepuberal students in gymnastics, namely students exercising squat vault, underswing, and handstand. To investigate cognitive processes of judgment formation, participants were asked to structure the movements via event segmentation as well as to explain their judgments. All teachers and trainers had experience in working with prepuberal children similar to those they observed in this experiment, and the teachers completed a gymnastics class during their studies.

Results: Judgment accuracy (with reference to judgments made by expert trainers) was found to be significantly lower in teachers compared to trainers ($p < 0.001$). Moreover, agreement on the ratings among teachers was lower than among trainers. Agreement about the temporal structuring of the tasks from event segmentation was lower among teachers than among trainers ($p < 0.05$). When explaining their ratings, trainers referred more often than teachers to kinematic features of the task that were relevant to the judgments.

Discussion: We discuss these findings in context of the teachers' task to perform accurate judgments. For suggestions on teacher training, we particularly emphasize the relevance of implementing knowledge about kinematic features of the tasks and student errors into real-life scenarios resembling the complex skill of making accurate judgments in the physical education classroom.

KEYWORDS

movement, middle school, competencies, feedback, learning, assessment, teacher education, sports

Highlights

- Teacher's judgment accuracy (with reference to judgments made by expert trainers) about movement errors in tasks in gymnastics is lower compared to the judgment accuracy of trainers.
- Agreement on the judgments among teachers is significantly lower than among trainers.
- Agreement on the temporal structure of the tasks is significantly lower in teachers than in trainers.
- Trainers referred more often than teachers to kinematic features of the task that were relevant to the judgments.

1. Introduction

Learning depends on accurate feedback. In school, diagnostic judgments by teachers are the primary source of information to generate feedback. Accordingly, accuracy of teachers' diagnostic judgments has been recognized and studied in various contexts (Südkamp et al., 2012; Loibl et al., 2020; Urhahne and Wijnia, 2021), such as in reading (Bates and Nettelbeck, 2001), mathematics (Leuders et al., 2022) and physical education (PE) (Ward et al., 2020; Moura et al., 2021). With regard to diagnostic judgments in PE, O'Brien et al., 2023 emphasized that "Physical education as a subject has evolved beyond the idea of 'busy, happy, good' (Placek, 1983), constituting a successful learning experience. Students developing and illustrating their capabilities across the cognitive and psychomotor domains are now at the forefront of physical education assessment (Hay, 2006)." The primary goal of assessment in PE classes is not on testing and grading (testing culture), but on the promotion of learning and teaching (i.e., assessment for learning) (López-Pastor et al., 2013; Tolgfors, 2018).

Development of motor competences is considered one of the pillars of PE (Sacko et al., 2021; Dudley et al., 2022), and external feedback about the students' performance is necessary for improving these competences (Magill, 2001; Leukel and Lundbye-Jensen, 2012). Feedback for learning requires that teachers make accurate judgments on characteristics critical to the performance of an intended skill (Sacko et al., 2021). One way to achieve this is by standardized testing (Seidel and Bös, 2012; Herrmann et al., 2016). However, standardized testing has significant limitations, namely: (i) the focus is typically narrow, meaning that tests capture only a small subset of motor abilities and skills. Thus, (ii) for many skills of the PE curriculum no test exists. Furthermore, (iii) ceiling and floor effects of tests are problematic because students at the extremes of the spectrum (high-ability and low-ability students) often cannot be reliably judged (Rink, 2013). The central deficiency of formal tests is that (iv) though they can produce accurate diagnostic judgments, they do not improve teachers' ability to judge students' performance. However, teachers' diagnostic judgments are a prerequisite for individual and adaptive feedback which supports learning (Swinnen, 1996; Magill, 2001; Leukel and Lundbye-Jensen, 2012). Therefore, diagnostic judgments as a basis for feedback should be integrated in PE lessons throughout and whenever possible. The ability to judge students on a continuous basis, which is considered a core component of teacher knowledge (Baumert and Kunter, 2013; Urhahne and Wijnia, 2021), cannot be substituted by a plethora of formal tests. Hence, besides formal testing teachers are required to make accurate judgments on characteristics critical to the performance of (complex) motor skills which are part of the PE curriculum, by focussing on key movement features that are relevant for performance of an intended task (Sacko et al., 2021). Analysing these characteristics is not trivial but requires both extensive practice in observation and profound knowledge of the intended task (Barrett, 1983; Ward et al., 2020).

There is evidence that PE teachers do not make accurate judgments (Lorente-Catalán and Kirk, 2016; van der Mars et al., 2018; Sacko et al., 2021), and reasons have been put forward trying to explain why this is the case, e.g., that criteria used for assessment are quite subjective and not based on evidence (Tolgfors, 2018). In fact, there is still very little information about the cognitive processes underlying judgment formation in PE teachers. This knowledge is crucial because it can substantially aid further research and interventions in teacher education (Loibl et al., 2020). It allows explaining the diagnostic skills of teachers (Chernikova et al., 2020; Loibl et al., 2020; Leuders and Loibl, 2021) and designing instructions by which teachers' diagnostic competences can be enhanced (Chernikova et al., 2020; Leuders et al., 2022). To exemplify this point, Niederkofler et al. (2018) found deficient judgment accuracy in teachers assessing fundamental motor skills and argued that they should be trained to become more competent in diagnostics. However, it remains unclear which aspects of the teachers' reasoning should be targeted, because the cognitive processes that lead to decisions were not evaluated in the study. Like Niederkofler et al. (2018), Ferrari et al. (2022) also found that teachers generally overrated the capabilities of their students on the whole-class level. They found significant correlations of judgment accuracy with class size, but not with the number of weekly lessons spent in the class, with experience, or self-reported competence. This may be attributed to the fact that judging the whole class requires the integration of individual judgments, which is more difficult for larger classes. However, for making such assumptions it is essential to gain more detailed knowledge about the information processing when making decisions.

Decision making, in PE and other domains, occurs through information processing which has recently been explicated within a theoretical framework (Loibl et al., 2020). In this framework, diagnostic thinking is conceptualized as three steps, namely perception, interpretation, and decision making. Perception is primarily visual when making judgments in PE. According to the mentioned theoretical framework (Loibl et al., 2020), it is crucial to consider (a) what kind of (visual) information the teacher perceives (referring to situational cues), and (b) how this relates to the decision (referring to the interpretation of the cues). In gymnastics, there is empirical evidence that judgments are based on visual perception and evaluation of salient kinematic features, like the height of an athlete over the ground when jumping over the vault, or the time airborne

(Takei, 1998; Farana and Vaverka, 2012; Luis del Campo and Espada Gracia, 2018; Mack, 2020). In general, and not constrained to gymnastics, experts were shown to be better in picking up and evaluating relevant perceptual cues compared to novices (Abernethy et al., 2001; Mann et al., 2007). According to these empirical findings, studying cognitive processes underlying judgment formation in teachers should therefore be concerned with kinematic features that teachers perceive and process. Experimentally, this can be achieved with eye-tracking (Kredel et al., 2017; Mack, 2020) or through event segmentation (Zacks and Swallow, 2007; Kurby and Zacks, 2008). Event segmentation is concerned with the idea of how people automatically and unintentionally compartmentalize perceptual experience into temporally defined phases that are segregated by event boundaries. A central postulate is that boundaries and phases are used by humans to make predictions and inferences (Zacks and Swallow, 2007). When observing movements, it has been shown that event boundaries relate to salient kinematic features (Zacks et al., 2009; Newberry et al., 2021). Event segmentation even works for movements characterized by rapid kinematic changes like tasks in gymnastics (Bläsing, 2015; Newberry et al., 2021; Stadler et al., 2021), in contrast to eye-tracking where short fixations of the eyes with a rapidly changing visual scene is difficult to track (Mack, 2020). Therefore, and according to the theoretical grounding about the relevance of kinematic features for decision making in gymnastics, event segmentation was used in the current study.

Diagnostic judgments in PE are volatile and often based on normative descriptions of correctness (Mechling and Munzert, 2004; Hong and Bartlett, 2008). Research on judgment accuracy in PE requires benchmarking teachers' achievements according to a reference source. Referencing can be done by comparing teachers' results to standardized test outcomes like in the study of Niederkofler et al. (2018). As mentioned, a disadvantage of test outcomes as reference is that for many tasks in PE no tests exist. An alternative approach, therefore, is to utilize trainers who hold a sufficient level of judgment accuracy with regard to the studied task as reference (Bläsing, 2015; Mack, 2020; Newberry et al., 2021; Stadler et al., 2021).

In summary, accurate diagnostic judgments on motor skills from teachers are important in PE. Empirically substantiated knowledge about cognitive processes of judgment formation in PE teachers is scarce but a requirement for understanding deficiencies in teachers' judgments and for furthering teacher education in this area. Therefore, in the present study, two aspects of diagnostic judgments, namely judgment accuracy and cognitive processes of judgment formation, were theoretically defined and empirically studied.

1.1. The present study

In the present study, we investigated PE teachers' diagnostic judgments of volatile and short-lasting tasks in gymnastics. This is typical for situations in the PE classroom, in that students sequentially perform the task, teachers observe this performance, and subsequently provide feedback on that performance. It is also typical for such a situation that teachers have to generate judgments and feedback under time pressure while teaching the whole class. The selected tasks, namely squat vault, underswing, and handstand, are part of the PE curriculum in Germany, and suitable for investigating information processing because the visual cues essential for estimating movement errors and performance are well-defined (Heinen, 2015). In order to draw meaningful comparisons, we included regular teachers and trainers in gymnastics who both had worked with prepuberal students before.

The study had two aims. The first aim was to explore to what extent teachers are able to form accurate judgments in the mentioned situations. We therefore investigated teachers' ratings on the severity of movement errors from watching video vignettes of prepuberal students performing squat vault, underswing, and handstand, and we compared these ratings to the ratings made by trainers. The video vignettes were played in real-time and pictured the students from the side, which resembles the situation that the teachers typically face in the PE classroom.

The second aim of the study was to investigate cognitive processes underlying judgment formation, which in our case relates to the processing of kinematic features linked to the decision about movement errors. This was achieved on the basis of event segmentation on the one hand and verbal reasoning on the other. With the former we assessed aspects of the visual information the subjects focussed on, and this was measured by spontaneous reactions of the teachers and trainers during the observation of the video vignettes. With verbal reasoning, we assessed the explicit reasons of a particular judgment, revealing the consciously driven process of the decisions. Accordingly, event segmentation informs about what kind of sensory cues the teachers and trainers actively focussed on, and the comparison between event segmentation and verbal reasoning reveals if these aspects were considered as meaningful for the decision by the teachers and trainers.

Although teachers received training on these three tasks (which are part of the basic repertoire in gymnastics) during their studies, and studied the characteristics of the movements before entering the experiment, we expected that they would have difficulties in identifying and interpreting relevant sensory cues for performance because of limited or absent practical experiences. In particular, the students gained mostly theoretical knowledge during their study but were not trained to apply this knowledge in a real-live scenario resembling the complex task of making accurate diagnostic judgments in the PE classroom. These difficulties should become apparent when processing the information (event segmentation and written explanations of the judgments) and also manifest in the accuracy of the judgments when compared to trainers.

2. Methodology and methods

2.1. Subjects

Forty subjects (aged between 21 and 60 years) participated in this study (Table 1). Half of the subjects (20 subjects, aged 29 ± 3 years, 13 males, 7 female) were trainee teachers in their final year of an 18 months induction phase ("Referendariat"). All of them were PE teachers, and all had worked for 1 year as PE teachers in secondary schools at the time of the experiment. Gymnastics had been part of their study. They all had taken a gymnastics class for one semester, and were trained on the tasks they had to judge in this study in terms of self-performance and teaching methods including movement characteristics and movement errors. Importantly, the knowledge they gained during their study had not been applied in a real-live task, thus

TABLE 1 Subjects' characteristics.

Trainers						
ID	Sex	Age	Licence as trainer	Licence as judge	Years active	Main sport(s)
1	m	52	A	A	45	Gymnastics
2	f	27	C		10	Gymnastics
3	m	27	A	B	15	Gymnastics
4	m	29	C		7	Gymnastics
5	m	59	C	B	30	Gymnastics
6	m	23	C		3	Gymnastics
7	m	54	B	B	20	Gymnastics
8	f	25	C		5	Gymnastics
9	f	25	C		4	Gymnastics
10	f	26	C	B	15	Gymnastics
11	m	29	C	B	12	Gymnastics
12	m	34	C		15	Gymnastics
13	m	49	B	A	40	Gymnastics
14	f	43	B	B	36	Gymnastics
15	f	25	C	C	17	Gymnastics
16	m	28	B	B	20	Gymnastics
17	m	22	C		15	Gymnastics
18	m	60	B	A	50	Gymnastics
19	m	21	B	B	13	Gymnastics
20	f	52	C		30	Gymnastics

Teachers						
ID	Sex	Age	Licence as trainer	Licence as judge	Years active	Main sport(s)
1	m	30				Handball, kitesurfing
2	m	26			15	Soccer, gymnastics
3	m	28				Karate, soccer
4	m	28				Snowboarding, soccer
5	m	34				Soccer
6	f	27				Athletics
7	m	27			17	Gymnastics
8	m	32				Volleyball, tennis
9	f	37				Volleyball
10	m	28				Handball, wakeboarding
11	m	28				Climbing, basketball
12	f	27				Running, swimming
13	f	28				Volleyball
14	f	28				Skiing, volleyball
15	m	28				Climbing
16	m	28				Soccer
17	m	36				Tennis, athletics
18	m	32				Handball
19	f	26				Volleyball, skiing
20	f	26				Soccer

The terms "licence as trainer" and "licence as judge" refers to the licence obtained as trainer and/or judge in gymnastics. m, male; f, female. Trainers marked in green were taken for creating the reference trainer (i.e., median ratings of the 10 best trainers). The reference trainer was used for calculating judgment accuracy of the teachers and the remaining 10 trainers.

they were not trained in performing accurate diagnostic judgments in situations similar to the situation in the classroom. Approximately 2 weeks before the experiment, the teachers were informed about the types of motor tasks and type of students they had to judge. When asked after the experiment, the teachers stated that they had prepared for it, by rehearsing the movement characteristics of the tasks by reading the scripts from their gymnastics class and/or reading about movement characteristics and movement errors of these tasks in (a) gymnastics book(s). The other half of the subjects (20 subjects, aged 36 ± 13 years, 13 males, 7 females) were gymnastics trainers holding a C-licence in gymnastics at minimum at the time of the experiment. All of the trainers spent between 3 and 8 h weekly in the gym and also trained children who were beginners in gymnastics. Gymnastics was their main sport. They had been active in gymnastics between 2 years and 15 years. Importantly, all subjects (teachers and trainers) worked with prepuberal children (teachers: school, trainers: gym) who achieved performance levels similar to the students they judged in the current study. All subjects (trainers and teachers) provided written informed consent before participation. The study was conducted according to the guidelines set in the Declaration of Helsinki (latest revision in Fortaleza) and approved by the local ethics committee. All subjects received a book voucher of 10 Euros to compensate for the time they spent in the laboratory.

2.2. Experimental design

Thirty video vignettes showing prepuberal female students performing gymnastics tasks were presented to the subjects. In each vignette, a single student was shown who performed a single task, namely a squat vault, an underswing, or a handstand with subsequent roll-out and stance (Figure 1). All three tasks are part of the curriculum for secondary schools in the State of Baden-Württemberg in Germany.

Subjects had to complete three different tasks in a consecutive order when watching the vignettes: first, they had to segment the video into meaningful, temporally defined phases. Second, the subjects had to rate the severity of movement errors of the overall performance of the student. Third, the subjects had to explain the main reasons for their rating in written form. These three tasks are explained in detail below.

Psychopy 3.0 (Open Science Tools Ltd.) running on a 13-inch Macbook Pro computer (5th Generation, Apple Inc., California) was used to control the execution and the timing of the tasks and to record the data. The laptop was connected to an external keyboard (Wireless Keyboard, 3rd Generation, Apple Inc., California) used by the subjects, and a 24-inch external LED screen (refresh rate: 60 Hz, LG, Seoul). The external screen was placed at a distance of 80 cm in front of the subjects.

2.3. Videos

Videos were recorded with a Sony 4 K camcorder (AX100 E, Sony, Tokio) from 10 female students aged between 10 and 12 years. The camera angle was chosen so that the shots captured the whole body of the students performing the tasks. There was no panning and zooming during the recordings. For handstand, the camera was placed at a distance of 3.5 m from and orthogonal to the mat on which the

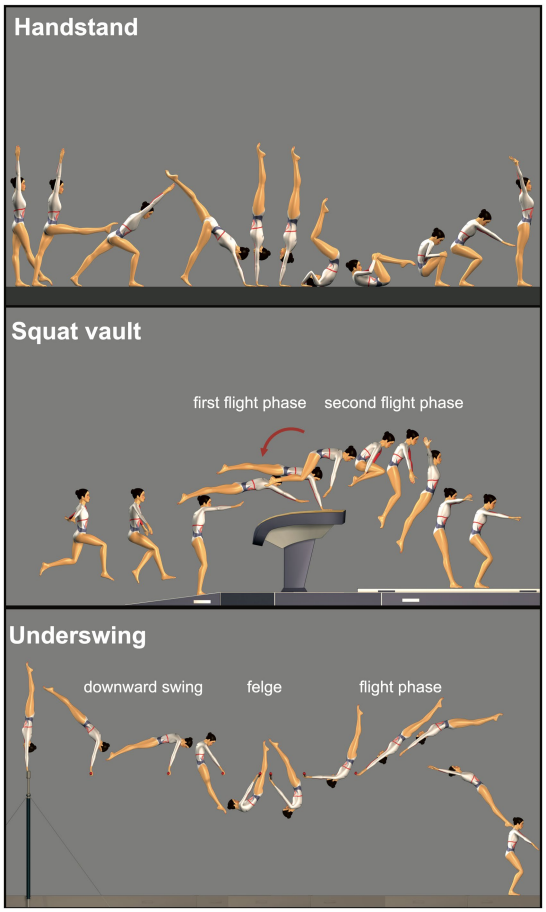

FIGURE 1
Illustrates movement phases for squat vault, underswing, and handstand. Note that for underswing, the athletes displayed in the videos in the current study did start with a support-position and not with a handstand, like it is displayed in the figure.

handstand was performed. For squat vault, the camera was placed at a distance of 5 m from and orthogonal to the vault. For underswing, the camera was placed at a distance of 4.8 m from the centre of the horizontal bar and orthogonal to the bar. For handstand and underswing, the camera captured the students from the beginning to the end of the task. For squat vault, the camera captured the students before jumping onto the springboard until landing on the mat after the vault. Thus, for squat vault, the camera did not capture the run-up. The contrast between the student and the background was increased by choosing light colors as background colors, i.e., yellow-colored vertically-placed mats when filming the underswing and the ivory-colored wall when filming the handstand and squat vault, respectively.

The raw video data were cut as follows: for handstand, the video started 1.5 s before the subjects initiated the movement, and ended after subjects reached stance. For squat vault, the video started 1.5 s before the subjects appeared in the picture (i.e., final step before the springboard), and ended after students reached stance (or crashed after landing). For underswing, the video started 1.5 s before the students started with the task, and ended after students reached stance (or crashed after landing). The final duration of each video was in between 4.2 and 6.8 s. The videos were presented in random order (i.e., no block design) to the subjects.

The students performing the tasks (age: 11–12 years) were recruited from a local gymnastics club and at the time of the recordings practiced gymnastics 1 to 2 times a week for a total of 2 h per week. The students had practiced this sport for half a year and up to 6 years at the time of the recordings. Thus, the performance level ranged between the children, which was intended, to cover the performance levels teachers typically see in the classroom (i.e., beginners to advanced). Each of the children performed all of the three mentioned tasks. All students wore the same clothes (namely black tight sports pants and a grey t-shirt) to reduce biased judgments relating to personal characteristics other than movement performance.

2.4. Event segmentation

Subjects had to segment the videos into temporally defined phases that were meaningful and seemed natural to them. Therefore, while watching the videos, they were instructed to press the spacebar on the keyboard whenever, in their opinion, a meaningful phase ended and a new one began. They had to place the index finger on the spacebar when performing the segmentation to reduce movement times affecting the timing of the presses. The time stamps marked by subjects are called event boundaries (Kurby and Zacks, 2008). Subjects viewed the entire video once before performing the segmentation. The segmentation had to be performed twice in consecutive order. Thus, subjects watched each of the videos three times (twice while also performing segmentation). When performing the segmentation the second time, we instructed the subjects to repeat what they did in the first run. Repetition was included because a previous study showed a systematic temporal shift of event boundaries with repeated exposure (Michelmann et al., 2021).

2.5. Judgment

Subjects had to rate the severity of movement errors they observed from viewing a task. They had to rate on a 5-point Likert scale: 1. no movement error; 2. task performed with minor movement errors; 3. task performed with medium movement errors; 4: task performed with major movement errors; or 5. task performed with very large movement errors. Note that subjects were asked about the severity of the errors, not the quantity (number) of errors they observed. The quantity indeed plays a role in competition but is not so important in a school setting. Here, effective performance-enhancive feedback rather addresses the severity of the error curtailing performance. Ratings had to be performed immediately after performing the segmentation. For each video, subjects were given 15 s to finalize their judgment, by pressing a number key on the keyboard referring to their rating (i.e., between 1 and 5).

2.6. Written explanations

After completing the judgment, subjects had to explain the main reasons for their rating in written form. They used the keyboard to write down the explanation in form of a text log. The subjects were instructed to explain the main reasons for the rating but were not constrained about what to include in and how to write the explanation (e.g., positive and negative aspects of the performance). They were

told to not pay attention to grammatical and language errors because spelling corrections were performed before analysing the data. For each video, subjects were given 50 s to write down the main reasons.

2.7. Choice reaction task

Subjects performed a choice-reaction task at the beginning of the experiment for assessing potential between-group differences in reaction times which could affect segmentation, namely how quickly subjects are able to press the button and set an event boundary. For the choice-reaction task, subjects viewed 5 different symbols (triangle, circle, cross, square, rectangle) on the computer screen (width: 10 cm, height: 10 cm, fill: light blue, background: white), which appeared in random order every 3.6 to 4 s for a duration of 300 ms. Each symbol was repeated 6 times; thus, 30 symbols were presented in total. A warning sign ("Get ready!") was presented on the screen for 4 s before showing the first symbol. Subjects were instructed to press the spacebar on the keyboard as quickly as possible as soon as a triangle appeared on the screen.

2.8. Experimental procedure

After having provided written informed consent, subjects were first tested in the choice-reaction task. Thereafter, they viewed a short video of 5 min in which the procedures of the main part of the experiment (i.e., event segmentation, rating, and explanation of the reasons for the rating) were explained. The subjects were allowed to ask questions concerning these procedures after having watched the instruction video. After the questions were answered by the experimenter the subjects executed three test trials and performed event segmentation, rating, and explanation of the reasons for the rating. The behavior of the subjects in these test trials was not recorded. The videos used for these test trials captured elementary school children from a local elementary school (fourth-graders) performing handstand, underswing, and squat vault, respectively. These videos were recorded and cut in the same way as the videos used in the main experiment. After finishing the test trials, subjects conducted the main experiment consisting of 30 videos. The duration of a single trial (event segmentation, rating, and explanation of the reasons for the rating for a single video) was 2 min and 10 s. The pause between two successive trials was 10 s. Thus, the overall duration of the main experiment was 1 h and 10 min. Subjects were given a break of 5 min after completing 15 videos to avoid fatigue. After completing 30 videos, at the end of the experimental session, subjects had to segment a final video showing a 10-year-old girl rising from a chair and leaving the room. This video showing a daily activity served as a control condition for the segmentation behavior. We expected that segmentation behavior would be different between trainers and teachers when viewing gymnastic tasks but not when viewing the daily activity.

2.9. Data analysis and statistics

2.9.1. Judgment

Interrater reliability within groups was assessed by calculating Krippendorff's alpha. Between-group differences (teachers versus trainers) of alpha were estimated by computing 95% confidence intervals from bootstrapping of the sample (2,000 sweeps) (Krippendorff, 2016). A significant difference between groups was assumed in case the confidence intervals of the groups did not overlap (Stolarova et al., 2014).

Rating accuracy was calculated with Spearman's rank correlations: the ratings (of 10 videos per task) of individual subjects were correlated with the ratings of a reference trainer. This approach of quantifying diagnostic accuracy via a correlation between the rank orders resulting from teachers' judgments on the one hand and from a reference order (often actual achievement in a test, but also expert judgments) is common in research concerned with diagnostic judgments. It goes back to a suggestion by Cronbach (1955), has been applied to teacher judgments by Helmke and Schrader (1987) and since then profusely and successfully used (Hoge and Coladarci, 1959; Südkamp et al., 2012; Urhahne and Wijnia, 2021). The reference trainer was created by calculating the median of the ratings per video of the 10 best trainers according to their experience level, namely their licence degree as trainer and judge in gymnastics (see Table 1). These 10 best trainers are assumed to perform the most accurate ratings. We compared (a) correlations between teachers and reference trainer (judgment accuracy of teachers) with (b) correlations between the remaining 10 trainers (those who were not taken for creating the reference trainer) and reference trainer (judgment accuracy of trainers). Statistical differences in judgment accuracy between groups (teachers against trainers) were assessed with unpaired Student's *t*-tests (for handstand, squat vault, and underswing, respectively) based on Fisher *z*-transformed correlation coefficients.

2.9.2. Segmentation behavior

Data from the segmentation task were aggregated in response vectors at a resolution of 60 Hz, corresponding to the screen refresh rate. Response vectors were set to 1 if a given participant had pressed the space bar within 200 milliseconds surrounding the time point and were set to 0 otherwise. To test for consensus between participants and the remaining members of the group (teachers and trainers) cosine similarity was computed between a participant's response vector and the average response vector across all other participants of the group. Accordingly, the average similarity to others' response was assessed per video. Differences in cosine similarity between groups were analysed with unpaired Student's t-tests.

The number of button presses was counted for each subject and video. Differences in button presses between groups were assessed with unpaired Student's *t*-tests.

Differences in the timings of button presses made between the first and the second run of segmentation were analysed by contrasting the instants of each of the button presses (with respect to video onset in deci-seconds) for the first and the second run. Values were discarded if the delay between run 1 and run 2 exceeded 500 ms, which was in most cases due to the fact that the subject did not perform a button press in run 1 or run 2, respectively. Linear regression was calculated from all button presses, for groups and tasks separately.

2.9.3. Segmentation behavior and movement characteristics

Movement characteristics linking to button presses were identified through analysing segmentation behavior in combination

with video analyses. The instants of the individual button presses (in deci-seconds from video onset) were marked in the videos, and movement characteristics occurring at these instants were identified. For squat vault and underswing, there were a small number of movement characteristics corresponding to these instants that were shared by members of the group and across videos, like the jump-off from the springboard for squat vault. According to the peaks of averaged response vectors from segmentation, 3 movement characteristics were selected for squat vault, and 4 movement characteristics were selected for underswing. The instants at which the movement characteristics occurred were determined for each video (in deci-seconds from video onset), and grand means were calculated from all of the 10 videos of each task.

In contrast to squat vault and underswing, for handstand the instants of button presses referred to a much larger number of movement characteristics, with larger inter-individual differences between subjects and videos. We therefore decided to select the top 5 characteristics shared by the trainers. Like for squat vault and underswing, the instants (in deci-seconds from video onset) at which these movement characteristics occurred were determined for each video, and grand means were calculated from all of the 10 videos.

2.9.4. Explanations

The 10 most frequently used nouns in explanations of the judgments were assessed separately for the two groups and the three tasks. The selection was performed according to the following procedure: first, the spelling of the written explanations was corrected. Second, the explanations of the subjects were tokenised and parsed into words. Capital letters were replaced by lower-case letters. Third, stop words were removed. Fourth, words with identical strings were counted. Fifth and finally, nouns were ranked according to the total count, and the 10 most frequently occurring nouns were listed.

2.9.5. Reaction times

Reaction times were recorded at a resolution of 60 Hz (screen refresh rate). Averaged reaction times were calculated for the 6 trials in each subject. Between-group differences in reaction times were analysed with an unpaired Student's t-test.

The level of significance was set to $p < 0.05$ for all tests. p-values from multiple comparisons were corrected according to Benjamini and Hochberg (1995). Data analyses were performed and graphs plotted using R programming language and R studio software (RStudio Inc., Boston).

3. Results

3.1. Judgment

This section of the results addresses the first aim of the study, which was to explore to what extent teachers are able to form accurate judgments. Therefore, teachers' ratings on the severity of movement errors were compared to the ratings made by trainers.

Single subject values of the ratings are depicted in Figure 2A. As it can be seen from Figure 2A, trainers declared movement errors to be more severe than the teachers across all tasks. The severity of movement errors was reported to be largest for underswing across groups, followed by squat vault and handstand.

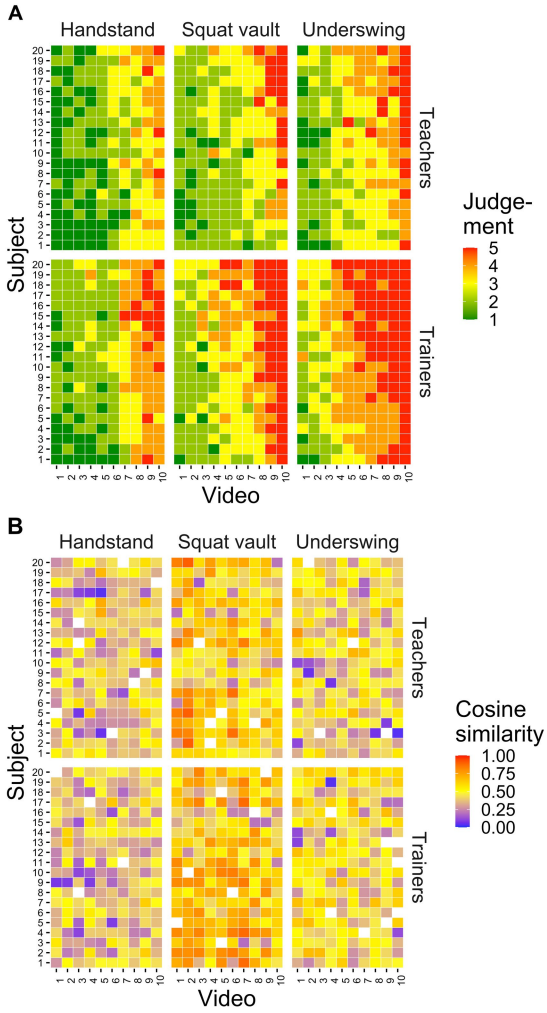

FIGURE 2
(A) Displays individual judgments. The dark green color indicates that, according to the subject's opinion, the task was performed with no movement error. The dark red color indicates that the task was performed with very large movement errors. The subjects were sorted according to the judgment mean (bottom: lowest value; top: highest value). Sorting was performed for each of the tasks separately. The same applies to the sorting of the videos. Videos were sorted according to the judgment mean from trainers (left: lowest value; right: highest value). **(B)** Displays cosine similarity calculated from segmentation data. Cosine similarity was computed between a participant's response vector (i.e., button presses of a single trial) and the average response vector (of the same trial) across all other participants of the group. Thus, the values display agreement between individual segmentation behavior and the segmentation behavior of others corresponding to the same group. A value of 1 indicates total agreement. Note that the sorting of the videos and subject IDs correspond to the sorting in part A of this figure. White spots: missing values, i.e., the subject did not perform segmentation.

Attributed performance levels clearly differed between students performing the tasks, ranging from students performing with no or few movement errors to students performing with very large movement errors. This was true for all three tasks.

Interrater reliability of the two groups and for the three tasks is displayed in Figure 3A. Krippendorff's alpha, expressing interrater reliability, was higher for trainers than for teachers for each of the

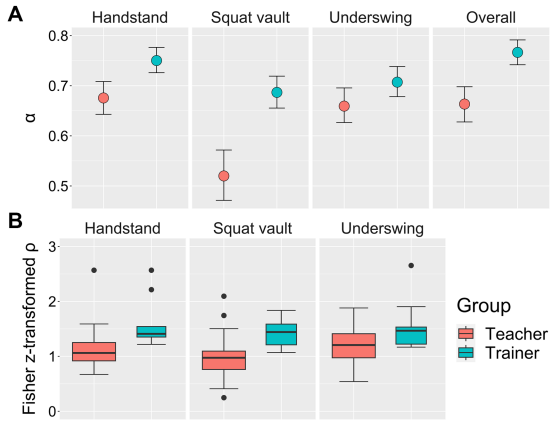

FIGURE 3
(A) Depicts Krippendorff's alpha calculated from judgments. "Overall" refers to the pooling of all data from the three tasks. Black bars display 95% confidence intervals which were calculated from bootstrapping of the sample (2,000 sweeps). **(B)** Depicts boxplots of Spearman's correlation coefficients (Fisher *z*-transformed values). Correlations were calculated between individual subjects and a reference trainer. Lower and upper hinges correspond to the first and third quartile, and the thick horizontal line of each boxplot represents the median. The upper whisker displays the largest value within 1.5 times the interquartile range above the third quartile. The lower whisker displays the smallest value within 1.5 times the interquartile range below the first quartile. Black dots display outliers. Note that, for reasons of clarity, one outlier in the trainers' group for handstand (*z*-score at 18.7) is not plotted.

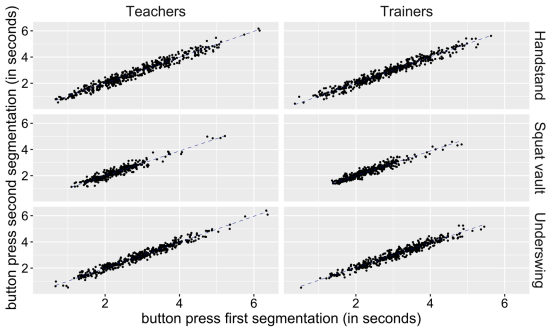

FIGURE 4
Shows correlations of individual button presses between the first and the second run of segmentation. The black dashed lines display the linear regression of data points.

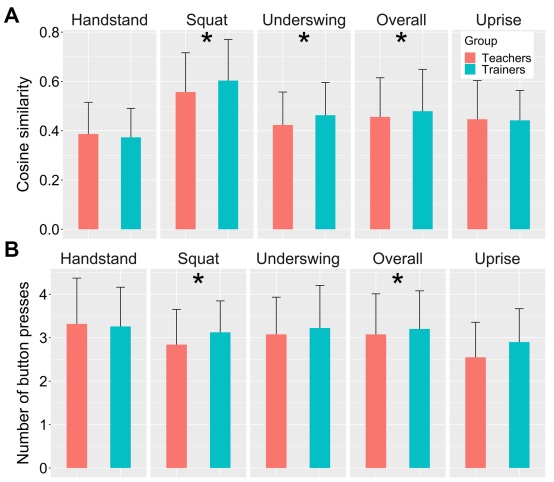

FIGURE 5
(A) Depicts grand mean values of cosine similarity from segmentation data. "Overall" refers to the pooling of all data from the three tasks. "Uprise" refers to the control condition (i.e., child stands up from a chair and leaves the room). Asterisks indicate significant differences between teachers and trainers. **(B)** Depicts grand mean values of the number of button-presses from segmentation. Asterisks indicate significant differences between the two groups.

three tasks and for the pooled data from all three tasks (Overall). 95% confidence intervals of alpha did not overlap between trainers and teachers for handstand, squat vault, and pooled data. This indicates significantly higher interrater agreement among trainers than among teachers.

The accuracy of the teachers' judgments was estimated with Fisher z-transformed Spearman's rank correlation coefficients that were compared between teachers and trainers (Figure 3B). Single subject values are depicted in Supplementary Figure S1. On the group level, results from the Student's *t*-tests yielded significant lower judgment accuracy in teachers compared to trainers for squat vault ($p < 0.01$, $t = -3.2$, Cohen's $d = 1.13$), underswing ($p < 0.05$, $t = -1.96$, Cohen's $d = 0.79$), but not for handstand ($p = 0.24$). Note that the result for handstand did not reach significance because of an outlier in the trainers' group (subject number 2) with a z-score at 18.7. Removing this outlier yielded a *p*-value of <0.05 ($t = -2.46$, Cohen's $d = 1.01$).

3.2. Segmentation

This section of the results addresses the second aim of the study, which was to investigate cognitive processes underlying judgment formation. Spontaneous reactions concerning the segmentation of the video vignettes were assessed and compared between trainers and teachers.

Timings of button presses in the first and second run of segmentation are displayed in Figure 4. Timings were not different between the two runs, as indicated by linear regression of data points. Slopes of all regressions were close to 1, and the regressions fitted the data points sufficiently well. Slopes (*S*) and coefficients of determination (R^2) were as follows: handstand: teachers ($S = 0.98$, $R^2 = 0.96$), trainers ($S = 1.0$, $R^2 = 0.96$); squat vault: teachers ($S = 0.95$, $R^2 = 0.92$), trainers ($S = 0.94$, $R^2 = 0.92$); underswing: teachers ($S = 0.99$, $R^2 = 0.96$), trainers ($S = 0.96$, $R^2 = 0.95$). According to these results, further analyses were conducted with data from the first run.

Single subject values are depicted in Figure 2B. As can be seen from Figure 2B, the biggest similarity of segmentation behavior among members of a group was observed for squat vault, followed by underswing and handstand, respectively.

There were no clear associations between cosine similarity and judgment. Declared severity of movement errors did not relate to the level of cosine similarity (comparison between Figures 2A,B). Otherwise, the coloring of the tiles in Figure 2B would follow the pattern of the coloring in Figure 2A. This is clearly not the case.

Cosine similarity was compared between the two groups for the three tasks, for the pooled data (overall), and for the control condition (uprise), respectively (Figure 5A). Unpaired Student's *t*-tests yielded

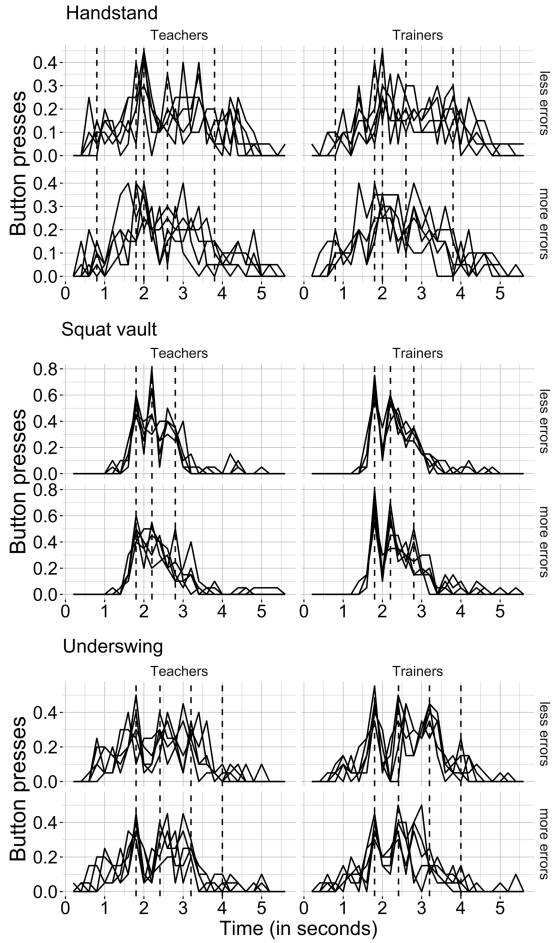

FIGURE 6
Displays averaged response vectors from segmentation over time, beginning at video onset. The 10 videos corresponding to each task were divided equally according to the judgment mean from trainers' ratings. Movement features that link to button presses from segmentation are displayed by dashed vertical lines. For handstand, the lines refer to: (i) subjects start moving, (ii) foot of supporting leg touches floor, (iii) hands touch floor (beginning of handstand), (iv) start of roll-out, (v) beginning of stance. For squat vault, the lines refer to: (i) feet leave springboard, (ii) hands leave vault, (iii) feet touch floor. For underswing, the lines refer to: (i) transition between backswing and forward swing, (ii) upper body crossing bar during forward swing, (iii) felge, (iv) feet touch floor. See also the text for a more detailed description of these movement features.

significant differences between teachers and trainers, for squat vault ($p < 0.01$, $t = -2.79$, Cohen's d = 0.28) and underswing ($p < 0.01$, $t = -2.92$, Cohen's d = 0.29). There were no differences for handstand ($p = 0.28$) and the control condition Uprise (i.e., child stands up from a chair and leaves the room) ($p = 0.92$). This indicates that, for squat vault and underswing, trainers' agreement about the temporal structuring of the tasks was significantly higher than the agreement among teachers.

Differences in the number of button-presses between groups are depicted in Figure 5B. For squat vault, the number of button presses was higher in trainers than in teachers (squat vault: $p < 0.01$, $t = -3.6$, Cohen's $d = 0.36$). There were no significant differences for handstand ($p = 0.56$), underswing ($p = 0.11$), and the control condition Uprise ($p = 0.17$). According to these results the higher number of button presses observed for the pooled data is likely due to the significant effect in squat vault.

For squat vault and underswing, button presses referred to distinct movement features marked in the average response vectors displayed in Figure 6. For handstand, in contrast to squat vault and underswing, there were no clear associations between response vectors and movement features across subjects and videos (see Figure 6). For squat vault and underswing, the movement features marked by the subjects refer to kinematic features that often occurred at transitions between subsequent movement phases. For squat vault, the first feature (in chronological order) marks a rapid change in acceleration of the body on the springboard, at the transition between the jump-off and the first flight phase. The second characteristic marks a rapid change in acceleration on the vault, at the transition between the first and the second flight phase (when the hands push off at the vault). The third feature marks the deceleration of the body at the transition between the second flight phase and the landing. For underswing, the first feature marks a rapid change in acceleration of the body at the transition between backswing and downward swing. The second feature marks the lifting of the body in the upward direction occurring between downward and upward swing. The third feature marks the maximum acceleration induced by the felge at the transition between the felge action and the flight phase. The fourth feature marks a deceleration of the body at the transition between the flight phase and the landing. Interestingly, for the underswing averaged response vectors of trainers indicated the presence/absence of the felge (cf. lower right part of Figure 6: the third dashed vertical line marks the transition between felge and flight phase). All subjects with fewer errors in Figure 6 (upper part) showed the felge, whereas all subjects with more movement errors (lower part) did not. Thus, trainers' segmentation behavior was sensitive to a critical feature for movement performance of the underswing.

3.3. Written explanations

In addition to segmentation, written explanations were analysed to investigate if the movement features marked when segmenting the videos were important for the judgment.

The 10 most frequently mentioned nouns of written explanations by trainers and teachers are listed in Figure 7. For squat vault, 5 out of the 10 nouns referred to movement features the trainers focussed on during segmentation, compared to 3 out of 10 in teachers. For underswing, 3 out of 10 nouns referred to features the trainers focussed on during segmentation, compared to 1 out of 10 in teachers. In total, 493 nouns referred to movement features the trainers focussed on during segmentation, compared to 236 nouns in teachers. This indicates that trainers referred to these features more often in their judgments (twice as much) compared to teachers.

4. Discussion

This study had two aims. The first aim was to elucidate to which extent teachers are able to perform accurate judgments. Therefore, we compared teachers' ratings to the ratings of trainers. The second aim was to investigate cognitive processes underlying judgment

Trainers

Handstand		Squat vault		Underswing	
handstand	153	**push-off (hands, vault)**	85	bar	84
roll-out	113	**start second flight phase**	67	**start forward swing (after backswing)**	75
arms/hands	86	**landing**	64	**felge/felgposition**	70
uprise	59	**start first flight phase**	61	hip	62
stiffness (body)	43	arms / hands	61	undershoot	49
back	30	vault	50	legs	40
head	28	**take-off (springboard)**	39	**landing**	32
feet	18	legs	37	shoulder	32
step (before handstand)	18	crash	34	height	30
upswing	17	touch down (on springboard)	29	tension	25

Teachers

Handstand		Squat vault		Underswing	
handstand	116	flight phase	96	sway	121
roll-out	111	**landing**	96	bar	76
arms/hands	80	arms/hands	87	legs	65
sway	64	**take-off (springboard)**	73	flight	50
stiffness (body)	51	vault	64	hip	47
uprise	41	legs	50	**landing**	47
back	41	sway	39	undershoot	46
legs	41	run-up	36	stiffness (body)	33
movement	33	spring board	22	feet	30
head	27	**stance**	20	body	29

FIGURE 7
Displays the most frequently mentioned nouns in written explanations of the judgments by trainers and teachers. The numbers indicate the total amount of entries in written explanations. Words in bold are expressions that refer to significant movement features from segmentation (see also Figure 2).

formation on the basis of event segmentation and written explanations of the judgments.

Concerning the first aim, we found significantly lower judgment accuracy in teachers compared to trainers. While this finding is in line with research on teachers' judgment accuracy in other domains (Südkamp et al., 2012), it is not trivial, given the fact that teachers received training on the tasks they judged and additionally prepared for the task they were required to accomplish, and the tasks are not difficult but part of the basic repertoire in gymnastics. Agreement on the ratings was significantly lower among teachers than among trainers. In general, teachers declared movement errors to be less severe compared to trainers.

Concerning the second aim, agreement about the temporal structuring of the tasks (squat vault and underswing) from event segmentation was significantly lower among teachers than among trainers. Trainers' segmentation behavior (i.e., button presses) referred to kinematic features that mostly indicated transitions between movement phases. Written responses from trainers, explaining the judgments, referred to these features more often than responses from trainers.

In the following, we discuss these results separately for each of the two aims, then discuss limitations of the study, and finally summarize the findings and their implications with regard to teacher education.

4.1. Teachers' ability to perform accurate judgments

There are two types of feedback students utilise when learning movements (Magill, 2001): intrinsic feedback refers to feedback that originates from the body's own sensors. Extrinsic feedback refers to feedback that originates from an external source, typically the teacher in a school setting. Accurate extrinsic feedback is necessary for motor learning of (complex) motor skills and for the consolidation of learned behavior (Leukel and Lundbye-Jensen, 2012; Leukel and Gollhofer, 2023). Accurate judgment is a prerequisite for providing accurate extrinsic feedback. In the current study, we observed lower judgment accuracy in teachers compared to trainers. Furthermore, ratings were less consistent among teachers than trainers. Indeed, Figure 7 visualizes the higher variability of teachers' ratings on individual students. What does this mean for PE? Are teachers' ratings deficient concerning their task to form accurate judgments in the classroom, and is this relevant? We argue that this is the case, and that it is relevant, and particularly refer to the inconsistency of judgments among teachers. Judgments on individual students did substantially vary between teachers, and because accurate judgment is a prerequisite for supportive feedback, a students would receive quite different feedback from different teachers. This raises concerns about the quality of extrinsic feedback necessary for learning (Magill, 2001; Leukel and Lundbye-Jensen, 2012; O'Brien et al., 2023). Furthermore, considering that the severity of movement errors may also be relevant for the grading, students would be graded quite differently from different teachers. Admittedly, PE covers a variety of tasks and not only tasks in gymnastics, and the goals are not just constrained to the development of physical competences. This means that not all judgments in PE have to be of high quality, nor can this be expected from teachers who typically spend only a fraction during their studies on gymnastics. However, we argue that if a task is considered important in PE, accurate judgment is a necessary ingredient for learning (Swinnen, 1996; Wolpert et al., 2011). We specifically consider the squat vault such a task because it is part of the basic repertoire in gymnastics, and considered relevant in PE and present in curricula from primary school to secondary school in many States in Germany.

Teachers' ratings are surely not as accurate as standardized kinematic measures. However, we refrain do conclude that teacher judgments should be replaced by standardized assessments tool (Seidel and Bös, 2012; Herrmann et al., 2016). We rather advocate that

teachers' judgment competences should be improved by training. As argued before, assessment tools are limited, with regard to focus (i.e., rather narrow, only a subset of motor abilities and skills are accessible), adaptability to individual needs, and ceiling and floor effects of the tests. Diagnostic competences of teachers are necessary because they overcome these limitations. Our results regarding the processes of judgment formation provide a starting point for elaborations on teacher training as will be discussed below.

We observed teachers to be milder in their judgments than trainers. Research on judgment accuracy in other subjects also shows that teachers usually overestimate their students (Ostermann et al., 2018; Oudman et al., 2018). This could be due to teachers' level of experience and/or due a personality trait. Teachers are less experienced than trainers with the tasks and thus might overlook movement errors. Concerning the personality trait, teachers might be generally milder in their judgments than trainers', because they not only focus on performance but also on social skills.

4.2. Cognitive processes of judgment formation

Investigating cognitive aspects of judgment formation is considered necessary for building a theoretical understanding of diagnostic judgments and deriving measures to improve diagnostic competences in teachers (Loibl et al., 2020). An important finding in the present study in this respect was that teachers had a less consistent concept of the temporal structuring of the tasks than trainers. Event boundaries, setting the temporal structure, have recently been shown to be important for memory formation and information retrieval (Michelmann et al., 2021). In the study of Michelmann et al. (2021), the time surrounding event boundaries was linked to information flows between cortex and hippocampus. The hippocampus is known for its role in memory formation and consolidation, and the area that is described as cortex in this study is linked to aspects of sensory (visual and auditory) integration and processing. According to these recently published findings, at event boundaries sensory inputs are likely compared to stored information. Thus, when judging tasks in gymnastics, at event boundaries the brain may perform comparisons of the target and the actual performance, separately for fundamental building blocks (i.e., different phases) of the observed act. This means that the brain may compare stored information of desired values of kinematic features with actual sensory (visual) values for each of the phases of the task separately, and the differences between desired and actual values indicate movement errors. It makes sense that this process does not cover the whole task but rather segregated parts, because this limits the amount of information that needs to be processed at once. Importantly, when trainers in the present study explained their ratings, the most frequently used nouns referenced to the kinematic features they focussed on during segmentation. This was also the case in teachers, but to a much lesser degree (half as much). This indicates that these features constitute the grounding of the ratings. Thus, according to these explanations, teachers may have more difficulties in (i) identifying kinematic features relevant for performance, and (ii) judging them appropriately.

Differences between trainers and teachers that concern the agreement of the temporal structuring were observed for squat vault and underswing, but not handstand. This could be due to the ambiguity of kinematic markers for handstand. For squat vault and underswing there are clear markers indicating salient changes in kinematics, like the feet leaving the springboard in squat vault indicating the beginning of the first flight phase. This might be the reason why, for handstand, event boundaries were set slightly differently from different subjects.

4.3. Limitations

This study has several methodological limitations: firstly, the button presses from segmentation do not clearly indicate to which aspects of the performance the subjects referred to. This is also due to a movement delay between the instant of the decision and the pressing of the button, and this delay varies between subjects (Norman and Komi, 1979; Kurz et al., 2019). The events in the videos that are referred to by the subjects are therefore not exactly traceable from the button presses.

Secondly, when analysing written explanations, we counted nouns/conjunct nouns and did not look for other word classes and combinations of other words except nouns. We did this on purpose because we were interested if subjects referred to movement features from event segmentation. In German language, movement features are expressed by (conjunct) nouns in combination with attributes (e.g., high *take-off*: hoher *Absprung*). It could well be that a thorough linguistic analysis would have brought up additional findings about the explanation of judgments. However, this was beyond the purpose of this study.

Thirdly, there is a risk that we might have overlooked semantically shared expressions pointing to movement features. This is particularly true for teachers. Teachers typically do not use technical lingo, in contrast to trainers (e.g., second flight phase in squat vault). They thus might have used various expressions referring to the same semantic content. We tried to account for this by searching for words with similar meanings, but there is still the risk that teachers used lingo we did not recognize as being similar.

Furthermore, there are some restrictions connected to more fundamental considerations: firstly, we cannot clearly define the level at which teachers' judgments are regarded as sufficient for students' development and grading. Statements about diagnostic competences in teachers were derived from comparisons to a reference group. Further empirical research is required to determine how the quality of a teachers' diagnostic feedback is coupled to students' learning.

Secondly, it is important to remember that accurate judgments are necessary for providing accurate feedback, but they are not sufficient. Improving teachers' diagnostic competences in the PE classroom does not necessarily mean that the students receive feedback that promotes learning. Providing feedback that promotes learning requires additional knowledge components (pedagogical content knowledge (Shulman, 1986), content knowledge, pedagogical knowledge) in addition to diagnostic information, which need to be flexibly applied according to the range of performance levels of the students (Hattie and Yates, 2014; Richartz et al., 2022). Indeed, a student who has difficulties to jump off the springboard requires very different feedback than a student who almost masters the squat vault.

Thirdly, the results of the present study are constrained to a limited number of tasks in PE, so findings only apply to a small portion within a broader range of tasks in sports. Hence, it cannot be concluded that PE teachers are poor judges in general.

Fourthly and finally, we did not measure pre-existing (declarative) knowledge about movement characteristics relevant for accurate

diagnoses in teachers. The extent of this knowledge might partly explain the interindividual differences we observed in the teachers.

4.4. Implications for teacher training

There is a broad consensus that diagnostics should be an integral part of PE with a strong aim to promote student learning and also teaching (Hay et al., 2015; O'Brien et al., 2023). Yet, teachers' practice does not meet this goal (Lorente-Catalán and Kirk, 2016; Moura et al., 2021). As a consequence, awareness of diagnostics in PE and promotion of diagnostic competences should be a fundamental part of teacher education (Ward et al., 2020; Moura et al., 2021; O'Brien et al., 2023). As mentioned before in this article, diagnostic judgments should not be constrained to standardized testing, but also need to be concerned with the assessment of (complex) motor skills which are part of the PE curriculum (Sacko et al., 2021). Accurate judgments on complex motor skills require knowledge about movement characteristics of the intended motor task, knowledge on typical student errors with respect to these characteristics, and extensive experience in observation (i.e., practical training in a real-life scenario) (Ward et al., 2020; Sacko et al., 2021). These aspects relate to the situational (cues) and person characteristics described in the framework of Loibl et al. (2020). According to the outcome of the present study, knowledge about movement characteristics and typical movement errors of the students are likely not sufficient for performing accurate judgments. Participants gained this knowledge and rehearsed it prior to the experiment but still showed insufficient performance. Performing judgments on motor skills in PE is a complex and a practical task involving many knowledge components which are declarative (e.g., knowledge about movement characteristics and typical movement errors) and procedural (e.g., knowledge about which position to take in the classroom for observing the students' performance, knowledge about where and when to look at the students while they perform the task) in nature. Accordingly, learning environments should acknowledge this complex nature of the task, including the procedural knowledge components. Instructional designs like the four-component instructional design model for training complex skills (4CID) (Van Merriënboer et al., 1992, 1997) do acknowledge these components, when they position the real-life scenario in the centre of the instructional design, and provide individual support in terms of knowledge and part-task practice. Hence, future studies may want to investigate how features of such an instructional design can effectively support the development of diagnostic competences in PE teachers.

5. Conclusion

In summary, in this study we found significant differences between trainers' and teachers' ratings on movement errors of tasks in gymnastics, with teachers' ratings being less accurate and consistent. The segmentation data indicated that the temporal structuring of the tasks was less consistent in teachers than trainers, and referred to kinematic features that are mostly linked to transitions between movement phases. In trainers, written explanations of the judgments contained these kinematic features from segmentation more often compared to teachers. We conclude from these results that diagnostic competences in teachers are insufficient and should be improved. According to the results from segmentation and written explanations, a preferable strategy for teacher education would be to focus on kinematic features relevant for performance in a practical teacher-training setting that resembles the real-life-scenario in the classroom, in which the different knowledge components (declarative and procedural) for making accurate judgments are integrated. Future studies might want to investigate the effectiveness of these learning environments on diagnostic competences in teachers.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Ethics committee at the University of Freiburg, Germany. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

CL, TL, FB, and KL: conceptualization, project administration, writing – review, and editing. CL: data acquisition. CL, TL, and KL: data analysis, interpretation of the data, and writing – original draft. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank Lars Breuning, Sabine Karoß, and Fabian Vogt, who helped with the recruitment of the subjects.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1162499/full#supplementary-material

References

Abernethy, B., Gill, D. P., Parks, S. L., and Packer, S. T. (2001). Expertise and the perception of kinematic and situational probability information. *Perception* 30, 233–252. doi: 10.1068/p2872

Barrett, K. R. (1983). A hypothetical model of observing as a teaching Skill1. *J. Teach. Phys. Educ.* 3, 22–31. doi: 10.1123/jtpe.3.1.22

Bates, C., and Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educ. Psychol.* 21, 177–187. doi: 10.1080/01443410020043878

Baumert, J., and Kunter, M. (2013). "The COACTIV model of teachers' professional competence" in *Cognitive activation in the mathematics classroom and professional competence of teachers*. (New York: Springer), 25–48.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bläsing, B. E. (2015). Segmentation of dance movement: effects of expertise, visual familiarity, motor experience and music. *Front. Psychol.* 5:1500. doi: 10.3389/fpsyg.2014.01500

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., and Fischer, F. (2020). Facilitating diagnostic competences in higher education—a meta-analysis in medical and teacher education. *Educ. Psychol. Rev.* 32, 157–196. doi: 10.1007/s10648-019-09492-2

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assumed similarity. *Psychol. Bull.* 52, 177–193. doi: 10.1037/h0044919

Dudley, D., Mackenzie, E., Van Bergen, P., Cairney, J., and Barnett, L. (2022). What drives quality physical education? A systematic review and meta-analysis of learning and development effects from physical education-based interventions. *Front. Psychol.* 13:799330. doi: 10.3389/fpsyg.2022.799330

Farana, R., and Vaverka, F. (2012). The effect of biomechanical variables on the assessment of vaulting in top-level artistic female gymnasts in world cup competitions. *Acta Gymnica* 42, 49–57. doi: 10.5507/ag.2012.012

Ferrari, I., Kühnis, J., Bretz, K., and Herrmann, C. (2022). "Diagnostische Kompetenz der Lehrpersonen und deren Bedeutung für die Förderung motorischer Basiskompetenzen", in *Narrative Zwischen Wissen Und Können*, Eds. Roland M. and Claus K (Baden-Baden: Academia), 175–194.

Hattie, J. A. C., and Yates, G. C. R. (2014). "Using feedback to promote learning", in *Applying science of learning in education: Infusing psychological science into the curriculum*. eds. V. A. Benassi, C. E. Overson and C. M. Hakala, Society for the Teaching of Psychology, 45–58., 45–58.

Hay, P. (2006). "Assessment for learning in physical education", in *The handbook of physical education*. eds. D. Kirk, D. Macdonald and M. O'Sullivan, (London, UK: SAGE), 312–25.

Hay, P., Tinning, R., and Engstrom, C. (2015). Assessment as pedagogy: a consideration of pedagogical work and the preparation of kinesiology professionals. *Phys. Educ. Sport Pedagog.* 20, 31–44. doi: 10.1080/17408989.2013.788145

Heinen, T. (2015). *Advances in visual perception research*. Milton Park, UKMilton Park Abingdon, Oxfordshire United Kingdom: Nova Science Publishers Inc.

Helmke, A., and Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teach. Teach. Educ.* 3, 91–98. doi: 10.1016/0742-051X(87)90010-2

Herrmann, C., Gerlach, E., and Seelig, H. (2016). Motorische Basiskompetenzen in der Grundschule. *Sportwissenschaft* 46, 60–73. doi: 10.1007/s12662-015-0378-8

Hoge, R. D., and Coladarci, T. (1959). Teacher-based judgments of academic achievement: a review of literature. *Rev. Educ. Res.* 59, 297–313. doi: 10.3102/00346543059003297

Hong, Y., and Bartlett, R. (2008). *Routledge handbook of biomechanics and human movement science* Routledge.

Kredel, R., Vater, C., Klostermann, A., and Hossner, E.-J. (2017). Eye-tracking technology and the dynamics of natural gaze behavior in sports: a systematic review of 40 years of research. *Front. Psychol.* 8:1845. doi: 10.3389/fpsyg.2017.01845

Krippendorff, K. (2016). *Bootstrapping distributions for Krippendorff's alpha*. Available at: https://www.asc.upenn.edu/sites/default/files/2021-03/Algorithm%20for%20Bootstrapping%20a%20Distribution%20of%20Alpha.pdf

Kurby, C. A., and Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends Cogn. Sci.* 12, 72–79. doi: 10.1016/j.tics.2007.11.004

Kurz, A., Xu, W., Wiegel, P., Leukel, C., and Baker, S. N. (2019). Non-invasive assessment of superficial and deep layer circuits in human motor cortex. *J. Physiol.* 597, 2975–2991. doi: 10.1113/JP277849

Leuders, T., and Loibl, K. (2021). Beyond subject specificity—student and teacher thinking as sources of specificity in teacher diagnostic judgments. *RISTAL* 4, 60–70. doi: 10.23770/RT1842

Leuders, T., Loibl, K., Sommerhoff, D., Herppich, S., and Praetorius, A.-K. (2022). Toward an overarching framework for systematizing research perspectives on diagnostic thinking and practice. *J. Math.-Didakt.* 43, 13–38. doi: 10.1007/s13138-022-00199-6

Leukel, C., and Gollhofer, A. (2023). Applying augmented feedback in basketball training facilitates improvements in jumping performance: augmented feedback in basketball. *Eur. J. Sport Sci.*, 1–16. doi: 10.1080/17461391.2022.2041732

Leukel, C., and Lundbye-Jensen, J. (2012). "The role of augmented feedback in human motor learning" in *Routledge handbook of motor control and motor learning*. eds. A. Gollhofer, W. Taube and J. B. Nielsen (Milton Park, Abingdon, Oxfordshire, United Kingdom: Routledge).

Loibl, K., Leuders, T., and Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (Dia CoM). *Teach. Teach. Educ.* 91:103059. doi: 10.1016/j.tate.2020.103059

López-Pastor, V. M., Kirk, D., Lorente-Catalán, E., Mac Phail, A., and Macdonald, D. (2013). Alternative assessment in physical education: a review of international literature. *Sport Educ. Soc.* 18, 57–76. doi: 10.1080/13573322.2012.713860

Lorente-Catalán, E., and Kirk, D. (2016). Student teachers' understanding and application of assessment for learning during a physical education teacher education course. *Eur. Phys. Educ. Rev.* 22, 65–81. doi: 10.1177/1356336X15590352

Luis del Campo, V., and Espada Gracia, I. (2018). Exploring visual patterns and judgments predicated on role specificity: case studies of expertise in gymnastics. *Curr. Psychol.* 37, 934–941. doi: 10.1007/s12144-017-9572-1

Mack, M. (2020). Exploring cognitive and perceptual judgment processes in gymnastics using essential kinematics information. *Adv. Cogn. Psychol.* 16, 34–44. doi: 10.5709/acp-0282-7

Magill, R. A. (2001). "Augmented feedback in motor skill acquisition" in *Handbook of sport psychology*. eds. R. N. Singer, H. A. Hausenblas and C. M. Janelle (Hoboken, New Jersey: John Wiley & Sons)

Mann, D. T., Williams, A. M., Ward, P., and Janelle, C. M. (2007). Perceptual-cognitive expertise in sport: a meta-analysis. *J. Sport Exerc. Psychol.* 29, 457–478. doi: 10.1123/jsep.29.4.457

Mechling, H., and Munzert, J. (2004). Handbuch Bewegungswissenschaft—Bewegungslehre. *Hofmann* 34, 231–236. doi: 10.1007/BF03176404

Michelmann, S., Price, A. R., Aubrey, B., Strauss, C. K., Doyle, W. K., Friedman, D., et al. (2021). Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nat. Commun.* 12, 1–15. doi: 10.1038/s41467-021-25376-y

Moura, A., Graça, A., Mac Phail, A., and Batista, P. (2021). Aligning the principles of assessment for learning to learning in physical education: a review of literature. *Phys. Educ. Sport Pedagog.* 26, 388–401. doi: 10.1080/17408989.2020.1834528

Newberry, K. M., Feller, D. P., and Bailey, H. R. (2021). Influences of domain knowledge on segmentation and memory. *Mem. Cogn.* 49, 660–674. doi: 10.3758/s13421-020-01118-1

Niederkofler, B., Herrmann, C., and Amesberger, G. (2018). Diagnosekompetenz von Sportlehrkräften—Semiformelle Diagnose von motorischen Basiskompetenzen. *Zeitschrift für sportpädagogische Forschung* 6, 72–96. doi: 10.5771/2196-5218-2018-2-72

Norman, R. W., and Komi, P. V. (1979). Electromechanical delay in skeletal muscle under normal movement conditions. *Acta Physiol. Scand.* 106, 241–248. doi: 10.1111/j.1748-1716.1979.tb06394.x

O'Brien, W., Philpott, C., Lester, D., Belton, S., Duncan, M. J., Donovan, B., et al. (2023). Motor competence assessment in physical education–convergent validity between fundamental movement skills and functional movement assessments in adolescence. *Phys. Educ. Sport Pedagog.* 28, 306–319. doi: 10.1080/17408989.2021.1990241

Ostermann, A., Leuders, T., and Nückles, M. (2018). Improving the judgment of task difficulties: prospective teachers' diagnostic competence in the area of functions and graphs. *J. Math. Teach. Educ.* 21, 579–605. doi: 10.1007/s10857-017-9369-z

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., and van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teach. Teach. Educ.* 76, 214–226. doi: 10.1016/j.tate.2018.02.007

Placek, J. H. (1983). Conceptions of success in teaching: busy, happy and good. *Teach. Phys. Educ.* 14, 46–56.

Richartz, A., Kohake, K., and Maier, J. (2022). "Pädagogische Qualität des Trainings im Kinder-und Jugendsport" in *Materialien für die Lehre im Gerätturnen: Bd. 3. Gerätturnen 2.0 (1. Aufl.). e-gymnastics*. ed. F. Bessi Germany: Freiburg.

Rink, J. E. (2013). Measuring teacher effectiveness in physical education. *Res. Q. Exerc. Sport* 84, 407–418. doi: 10.1080/02701367.2013.844018

Sacko, R. S., Utesch, T., Cordovil, R., De Meester, A., Ferkel, R., True, L., et al. (2021). Developmental sequences for observing and assessing forceful kicking. *Eur. Phys. Educ. Rev.* 27, 493–511. doi: 10.1177/1356336X20962134

Seidel, I., and Bös, K. (2012). Chancen und Nutzen motorischer Diagnostik im Schulsport am Beispiel des DMT 6–18 Sportunterricht, Schorndorf, 61, 228–233.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14. doi: 10.3102/0013189X015002004

Stadler, W., Kraft, V. S., Beer, R., Hermsdörfer, J., and Ishihara, M. (2021). Shared representations in athletes: segmenting action sequences from taekwondo reveals implicit agreement. *Front. Psychol.* 12, –733896. doi: 10.3389/fpsyg.2021.733896

Stolarova, M., Wolf, C., Rinker, T., and Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Front. Psychol.* 5:509. doi: 10.3389/fpsyg.2014.00509

Südkamp, A., Kaiser, J., and Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *J. Educ. Psychol.* 104, 743–762. doi: 10.1037/a0027627

Swinnen, S. P. (1996). "Information feedback for motor skill learning: a review" in *Advances in motor learning and control*. ed. N. Zelaznik (Champaign, IL 61820: Human Kinetics), 37–66.

Takei, Y. (1998). Three-dimensional analysis of handspring with full turn vault: deterministic model, coaches' beliefs, and judges' scores. *J. Appl. Biomech.* 14, 190–210. doi: 10.1123/jab.14.2.190

Tolgfors, B. (2018). Different versions of assessment for learning in the subject of physical education. *Phys. Educ. Sport Pedagog.* 23, 311–327. doi: 10.1080/17408989.2018.1429589

Urhahne, D., and Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educ. Res. Rev.* 32:100374. doi: 10.1016/j.edurev.2020.100374

van der Mars, H., McNamee, J., and Timken, G. (2018). Physical education meets teacher evaluation: supporting physical educators in formal assessment of student learning outcomes. *Phys. Educ.* 75, 582–616. doi: 10.18666/TPE-2018-V75-I4-8471

Van Merriënboer, J. J. (1997). *Training complex cognitive skills: a four-component instructional design model for technical training*. Englewood Cliffs, NJ: Educational Technology.

Van Merriënboer, J. J., Jelsma, O., and Paas, F. G. (1992). Training for reflective expertise: a four-component instructional design model for complex cognitive skills. *Educ. Technol. Res. Dev.* 40, 23–43. doi: 10.1007/BF02297047

Ward, P., Ayvazo, S., Dervent, F., Iserbyt, P., Kim, I., and Li, W. (2020). Skill analysis for teachers: considerations for physical education teacher education. *J. Phys. Educ. Recr. Dance* 92, 15–21. doi: 10.1080/07303084.2020.1853635

Wolpert, D. M., Diedrichsen, J., and Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nat. Rev. Neurosci.* 12, 739–751. doi: 10.1038/nrn3112

Zacks, J. M., Kumar, S., Abrams, R. A., and Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition* 112, 201–216. doi: 10.1016/j.cognition.2009.03.007

Zacks, J. M., and Swallow, K. M. (2007). Event segmentation. *Curr. Dir. Psychol. Sci.* 16, 80–84. doi: 10.1111/j.1467-8721.2007.00480.x