



Why assessment in physical education is still problematic: A critical interpretive synthesis of physical education assessment literature

Dean Barker^{1,2} , Björn Tolgfors¹ , and Annica Caldeborg¹ 

Abstract

Assessment has received considerable attention from researchers in the field of physical education (PE). Many scholars have examined either formative assessment or summative assessment, with their focus leading to different questions and considerations. In this review, we examine how and why both formative and summative assessment have been problematized by PE scholars. Through a critical interpretive synthesis, we identify: (1) the main problems associated with both forms of assessment identified between 1999 and 2024, and (2) the solutions that scholars have offered in response to these problems. Problems with summative assessment center on teachers' use of personal and internalized criteria, students' negative experiences, and the guidance that policy provides teachers for enacting assessment. Solutions revolve around the provision of continuing professional development, improving initial teacher education, and ensuring that policy clearly delineates how assessment should be conducted. Problems with formative assessment revolve around teachers' and students' unfamiliarity with formative assessment practices and their lack of competence in using assessment strategies. Recommended solutions center on accepting that formative assessment has advantages and disadvantages, increasing students' participation in assessment practices, and improving teachers' assessment proficiency. We consider the extent to which assessment scholarship can contribute to change in assessment practices in PE, developing the thesis that several factors constrain the ability of research to lead to improvements. We conclude with alternative approaches that scholars might use to reimagine research on formative and summative assessment.

¹University of Örebro, Sweden

²Oslo Metropolitan University, Norway

Corresponding author:

Dean Barker, School of Health Sciences, University of Örebro, Fakultetsgatan 1, SE701-82 Örebro, Sweden.

Email: dean.barker@oru.se

Keywords

Assessment for learning, assessment for accountability, formative assessment, summative assessment, instructional alignment

Introduction

Assessment has received considerable attention from researchers in the field of physical education (PE). In the last three decades, scholars have presented philosophical justifications for assessment (Penney et al., 2009; Rink and Mitchell, 2002; Starck et al., 2018), proposed concrete strategies for educators (MacPhail et al., 2023; Melograno, 1998; Scanlon et al., 2023), outlined conditions necessary for its successful implementation (Hay and Penney, 2009), issued position statements on best assessment practices (AIESEP, 2020), and critiqued those same position statements (Tolgfors and Barker, 2023). The result has been a burgeoning corpus of research addressing multiple philosophical, theoretical and practical aspects of assessment. A broad survey of this body of work, however, reveals two related and potentially problematic patterns. One is a recent tendency to devote more attention to formative assessment than summative assessment. Five peer-reviewed literature reviews have, for instance, focused on formative assessment (Bores-García et al., 2020; Herrero-González et al., 2024; López-Pastor et al., 2013; Moura et al., 2021; Otero-Saborido et al., 2021), while none, to our knowledge, have focused on summative assessment. A second pattern concerns the way formative and summative assessment have been presented in scholarship. Formative strategies have been described as *meaningful*, *relevant*, *learning-focused*, *authentic* and *democratic* (Bores-García et al., 2020; Chng and Lund, 2018) and criticism of formative assessment principles has been relatively rare (DinanThompson and Penney, 2015; Starck et al., 2018). In contrast, problems identified with both the principles and the practices of summative assessment have appeared frequently in PE research. The idea that teachers assess their students based on personal or internalized criteria instead of criteria set out in curricula has, for example, resurfaced many times over the years (AIESEP, 2020; Hay and Penney, 2013; Matanin and Tannehill, 1994) with little indication of the problem being resolved.

In this review, we examine how and why both formative and summative assessment have proven problematic in PE. Through a critical interpretive synthesis (CIS) (Dixon-Woods et al., 2006), we identify: (1) the main challenges, difficulties, problems, tensions, and disadvantages (referred to from here simply as “problems”) associated with forms of assessment identified between 1999 and 2024, and (2) the solutions that scholars have offered in response to these problems. We then consider whether the recommendations offered by scholars have the potential to change assessment practices in PE. We conclude by discussing alternative approaches that researchers might use to reimagine research on formative and summative assessment.

Background

Following Hay and Penney (2013: 6), we use the term *assessment* to refer to “any action of information collection within education settings that is initiated for the purpose of making some interpretive judgements about students.” Acknowledging considerable variation in terminology in the general education assessment literature (AIESEP, 2020; Black and Wiliam, 2009), we distinguish between formative assessment practices, where data is collected with an intention to inform, support and promote student learning (also referred to as *assessment for learning*, see Black and Wiliam,

2009), and summative assessment practices, where data is collected with an intention to provide an account of learning that has taken place (sometimes referred to as assessment *of* learning). In practice, assessment tasks are rarely intended to be only formative or only summative, and the intentions behind assessment tasks vary in degree rather than in kind (AIESEP, 2020). The PE literature, however, tends to distinguish between two “forms” of assessment and this distinction is therefore reflected in our handling of the material.

López-Pastor et al. (2013) traced PE scholars’ growing interest in formative assessment back to the 1980s when physical educators—largely concerned with improving pupils’ motor skills and fitness—recognized that tests of isolated skills and fitness components said little about what pupils could do beyond the particular test. Dissatisfaction with traditional forms of summative assessment, along with curricular developments in PE (O’Sullivan, 2013) and the growing influence of constructivist theories of learning more generally (Hay and Penney, 2013), provided impetus for physical educators to develop new assessment approaches. Scholars became concerned with assessment strategies that could promote student learning, that were explicitly tied to students’ lives, and that could be used to address content beyond motor skills and fitness levels (Grehaigine et al., 1997; Hay and Penney, 2009; Oslin et al., 1998; Siedentop and Tannehill, 2000).

Since the 1980s, formative assessment has gained acceptance in PE and has been the focus of a number of scientific investigations (Bores-García et al., 2020; Herrero-González et al., 2024; López-Pastor et al., 2013; Moura et al., 2021; Otero-Saborido et al., 2021). Recent reviews provide useful information on the locations of the studies being conducted, the educational stages investigated, the types of formative assessment being enacted, and the content being addressed—see for example, Bores-García et al.’s (2020) review on peer assessment (13 articles) and Otero-Saborido et al.’s (2021) review on self-assessment (13 articles). López-Pastor et al.’s (2013) synthesis of 51 articles adds a valuable historical contextualization of the topic, as well as a comparison of alternative and traditional forms of assessment. López-Pastor et al.’s (2013) work has since been updated by Moura et al. (2021), who reviewed 52 texts with the aim of ascertaining the prevalence of formative assessment in schools. Common across these reviews is an assumption that formative assessment is beneficial and that physical educators should make use of formative assessment strategies. If scholars have critiqued formative assessment, it has been to either highlight practical difficulties or lament its absence in schools (see also DinanThompson and Penney, 2015; Herrero-González et al., 2024; Starck et al., 2018).

Summative assessment has received less attention than formative assessment in recent times. Some scholars have suggested that accountability-oriented assessment is necessary for securing the legitimacy of PE as a school subject (Collier, 2011; Sundaresan et al., 2017). Questions of legitimacy have a long history in PE scholarship (Matanin and Tannehill, 1994). More than two decades ago, Rink and Mitchell (2002: 206) claimed that PE needed to fall into line with educational developments that privileged “standards, assessment and accountability” if it was to be seen as a legitimate subject on the school curriculum. Presaging an emphasis on instructional alignment that has since pervaded much assessment research (MacPhail et al., 2023; Scanlon et al., 2023), Rink and Mitchell (2002: 209) proposed that teachers needed to “clearly articulate what students are expected to learn, have assessment materials to determine if students did learn, and have a mechanism for assessing large numbers of students and schools.”

Generally accepting the importance of summative assessment for both legitimacy and student learning purposes, scholars have since claimed that enacting summative practices in PE is challenging and not without risks. Thorburn (2007), for example, offered a detailed analysis of how educational traditions that divide theory and practice make summative assessment exigent for PE teachers. Looking closely at the consequences of assessment for students in PE, Hay and Penney

(2013) maintained that an emphasis on summative assessment can result in students ignoring substantial learning and focusing instead on achieving grades (see also Tolgfors and Barker, 2023).¹ Drawing on Ball's (2003) notion of performativity, Hay and Penney (2013: 8) suggested further that summative assessment can also place teachers under "inordinate professional pressure ... to achieve high cohort performance standards." In Hay and Penney's (2013) view, it is necessary for educators to carefully distinguish between summative assessment, which is the process whereby data is collected and used to provide an account of student learning, and grading, which is typically a small part of the summative process.

In short, researchers interested in assessment in PE have broadly advocated for formative assessment strategies and have been more ready to critique summative assessment strategies. In the absence of a concerted examination of the problems associated with both formative and summative assessment across PE scholarship, this article provides a critical investigation of the problems that arise, and the solutions scholars have offered in response to these problems. In the next section, we outline our methodological approach to the review.

Methodological approach: CIS

CIS is a review methodology that involves the synthesis of large volumes of diverse data (Farias and Laliberte Rudman, 2016). CIS draws on the tradition of qualitative inquiry, emphasizing a *critical orientation*, *theory development*, and *flexibility* (Dixon-Woods et al., 2006). Our synthesis was critically oriented in that it comprised an examination of the underlying assumptions and values (Farias and Laliberte-Rudman, 2016) that underpin research on assessment in PE. In terms of theory development, our synthesis involved inductively developing a synthesizing argument from the categorization of problems and solutions in the research. Our argument integrates evidence from across studies into a coherent thesis (Dixon-Woods et al., 2006: 5), specifically seeking to explain why problems with assessment persist. Finally, flexibility is evident in our iterative development of research questions and our search and selection processes (see below). We describe our procedure under six "activity" headings (Depraetere et al., 2021). These activities did not proceed in discrete stages. Rather, each activity overlapped, and experiences during one activity helped us to develop the other activities (Farias and Laliberte-Rudman, 2016).

Formulating open research questions

After surveying reviews of literature and conceptual texts (i.e. the literature presented in the Introduction and Background sections above), we developed two "compass questions" (Bullock et al., 2021): (1) How does assessment become problematic in PE? and (2) What solutions do PE scholars offer for addressing the problems they identify? These questions allowed us to narrow our focus in our search for literature and guided our extraction of data once we had found the literature for review.

Searching for literature with a broad strategy

We adopted two search strategies to identify relevant literature. First, we identified texts using the databases Google Scholar and Education Resources Information Center (ERIC). The search combinations employed were "physical education" AND "assessment" and "PE" AND "assessment," which we used "anywhere in the article." Second, we relied on snowball retrieval, where we searched reference lists from the texts found with the first strategy to identify further texts.

In line with CIS methodology (Dixon-Woods et al., 2006), our intention with this broad search was to find potentially relevant literature from which we could make a more precise selection based on relevance.

Selecting literature based on relevance

To be included in the review, texts needed to: (1) Have been peer-reviewed and published in English; and (2) Have been published between 1999 and 2024. The start year reflects our recognition that both assessment and PE have been affected by societal changes (e.g. increased use of digital technologies) since the turn of the millennium. Our intention with the time span was to be comprehensive but current; (3) Refer to problems, difficulties, or disadvantages with enacting formative and/or summative assessment in school PE. These problems could be described from teachers', preservice teachers', students', researchers' or other stakeholders' points of view. Problems did not need to constitute the analytic focus of the study, but assessment did. In other words, if the text referred to assessment but focused on physical literacy (e.g. Jean de Dieu and Zhou, 2021) or fundamental movement skills (Lander et al., 2015) or teachers' content knowledge (e.g. Ward et al., 2015), for example, the article was not included; (4) Describe a study that involved producing empirical material. Employing this last criterion was based on our intention to contextualize the problems identified.

Evaluating the quality of the literature

We used Dixon-Woods et al.'s (2006: 4) "appraisal prompts for informing judgements about the quality of papers" (see the Appendix) to decide whether a text should be included in the review. In short, if we had concerns regarding the clarity or appropriateness of the study's aims, objectives, research design, methodological approach, and/or interpretation, we sent the text to the other two authors for cross-checking. If, after discussion, we agreed that parts of the text were unclear or aspects of the investigation had been conducted inappropriately, we deemed it to be of insufficient quality and excluded the text from our sample (see Figure 1 for a summary of the search and selection process). Details of the 48 reviewed texts are included in the supplemental material available online. They have been assigned a number (1–48) and are referenced in this article using that number.

Extracting data from the literature

Unlike in traditional systematic reviews (Bores-García et al., 2020; Otero-Saborido et al., 2021), we were interested in extracting a relatively narrow range of data from the literature, namely the type of assessment in focus, the country, the school level, the reasons for how and why formative and summative assessments become problematic in PE, and the solutions that scholars recommended. To extract this information, we developed a template that we used to record details from each article (see Table 1). Relevant parts of the text were either paraphrased or copy-pasted into the template so that each article had its own Word document. Following Depraetere et al. (2021), we initially systematized the process by completing data extraction individually with the same five articles and then compared our results. Reaching an agreement on the type of assessment in focus, country, and school level was unproblematic. Questions did arise around the identification of problems and solutions so in line with other CIS research (Dixon-Woods et al., 2006), this aspect of data extraction was completed in an inclusive manner. This meant that if there was any question of whether text from the article should be interpreted as describing a problem or solution, it was

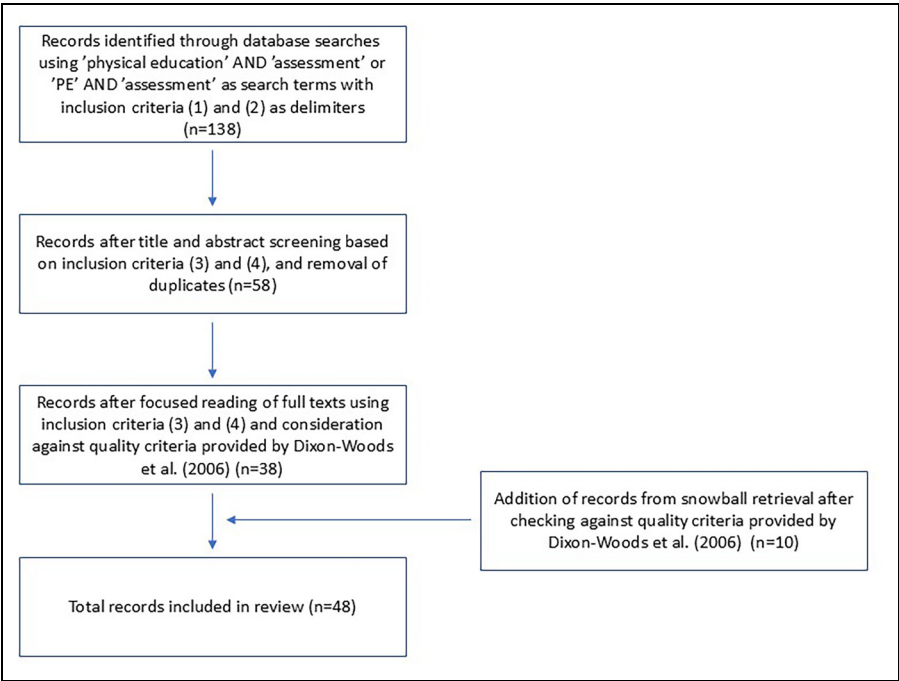


Figure 1. Flowchart of literature search and selection process.

Table 1. Extraction template.

Type of assessment in focus, including year level.	Type of study conducted, including country, number of participants and methodology.	Problem(s) identified— explanation of how and why assessment becomes problematic.	Solutions recommended.
--	---	---	------------------------

included and compared against other completed templates later in the process. When templates for all studies were completed, the first author printed all templates, grouped them by assessment type (formative, summative or “formative and summative”), and compared the content of the templates for patterns and similarities. Using pen and paper, the first author categorized problems and solutions under tentative headings such as “teachers’ personal/internalized judgments” and “students’ inexperience.” The headings and placement of example text under the headings were discussed and finalized at a data analysis meeting.

Formulating a synthesizing argument

The final activity entailed creating a synthesizing argument (Farias and Laliberte-Rudman, 2016). Formulating this argument involved critically examining the summary of the extracted data and considering the assumptions that make the problems identified persistent and the solutions ineffective or difficult to enact in practice. The first author started the formulation process by constructing a draft thesis in response to the questions above. Once this draft had been completed, the second and

third authors refined the text, adding, deleting and commenting where they saw necessary based on their familiarity with the literature (see Bullock et al., 2021). In this way, the thesis took form as a synthesis of the problems and potential solutions associated with assessment in PE.

Findings

Of the 48 articles, 19 dealt with formative assessment, 23 dealt with summative assessment, and six dealt with both formative and summative assessment. Eight focused on primary/elementary levels, 34 focused on secondary or upper secondary levels, and six focused on primary, secondary and upper secondary levels.² We have chosen not to distinguish between levels in our synthesis because the format that PE took and the content that PE teachers dealt with in the various contexts were not always clear. It was impossible to say, for example, whether PE involved classroom-based lessons at secondary or upper secondary levels in all countries. Many of the articles came from Scandinavia (21), followed by North America (9) and the United Kingdom (6). Specific characteristics of each reviewed text can be found in the supplemental material. The remainder of the section focuses on problems and solutions, associated first with summative assessment and then with formative assessment. The findings are summarized in Table 2.

Problems with summative assessment

Problems identified with summative assessment can be divided into three broad “areas,” each of which contains related subproblems. These problematic areas overlap but cluster around: (1) the intrusion of teachers’ personal/internalized criteria into the assessment process, (2) students’ negative experiences of assessment, and (3) a lack of curricular clarity regarding how assessment should be conducted.

Teacher use of personal criteria is framed as a problem in that PE teachers collect data for the purposes of providing accounts of student learning without basing the data on criteria that reflect learning (11, 37). Many investigations suggest that teachers use behavioral, dispositional and/or physical factors to assess their students. These factors include whether students: have the correct clothing and participate in lessons (11, 24, 47); are fit and/or appear in good shape (33, 34); are willing to try new activities (39, 46); and demonstrate competitiveness or aggression during games (31). Some studies further suggest that teachers are unaware of the impact their internalized criteria have on their assessment practices (2, 39). Beyond the contradiction of using criteria that do not reflect learning to create accounts of learning—a contradiction that calls into question whether PE teachers are engaging in summative assessment *at all* when they rely on such criteria—the reliance on personal criteria has been problematized in three ways. Scholars have observed that personal criteria fall outside curricular prescriptions (7), thus creating instructional misalignment between learning goals and assessment practices (11, 26). Further, scholars have pointed out that assessments based on teachers’ internalized criteria lack transparency. Several investigations highlight the gap between the criteria teachers use and the criteria on which students believe they are being assessed (27, 31, 48). Lastly, some scholars have noted that when teachers use their own subjective criteria, it becomes difficult to compare what students have learned across schools or even across classes at the same school (17).

The second problematic area concerns students’ negative experiences of assessment. Negative experiences comprise experiences of assessment as unfair and experiences of assessment as anxiety-inducing. In Røset et al.’s (2023) (34) investigation, boys and girls both suggested that the opposite sex were advantaged in PE assessment (see also 23 for statistics regarding gender

Table 2. Summary of findings.

Summative assessment	
Problems	Solutions
Intrusion of teachers' personal performance criteria, where teachers base their assessment on behavioral (e.g. punctuality), dispositional (e.g. competitiveness) and physical (e.g. appear to be in good shape) factors. The use of personal or internal criteria results in: – A lack of instructional alignment, since these criteria are not set out in curricular documentation. – A lack of transparency, since teachers do not make explicit the factors they are using to make the assessment. – A lack of comparability, since students in different classes or schools cannot be compared because their teachers are basing their assessment on different factors. Students experience summative assessment as: – Unfair or inequitable because they perceive assessment structures as advantaging certain groups of students, for example, boys, students with experience of sport. – Stressful because they believe their assessment performance is open to scrutiny from their peers. Official curricular statements are either too narrow, leaving teachers little room to use their professional judgment, or they are too vague and do not make clear for teachers how assessment should be conducted.	Improve initial teacher education and/or provide continuing professional development to help teachers: – Become more reflective regarding how their personal preferences affect their grading practices. – Gain a better understanding of official curricular statements. – Learn to align assessment with teaching practices and learning intentions. – Become more proficient at communicating learning intentions, collecting evidence, and evaluating this evidence. Ensure that assessment policy contains support material that can guide and inform teachers' assessment practices.
Formative assessment	
Problems	Solutions
Teachers are unfamiliar with, and lack competence in, enacting formative assessment, including developing learning targets and providing feedback. Teachers express concerns when attempting to enact formative assessment including: – Balancing time needed for assessment with time needed for other tasks and activities. – Keeping control in the classroom while encouraging students to make decisions. Students lack experience and proficiency with formative assessment strategies such as giving and receiving feedback. This problem may be exacerbated by students' status and power relations.	Accept that there are different ways of doing formative assessment and that each way has advantages and disadvantages. Improve teachers' proficiency of working with formative assessment through initial teacher education and continuing professional development. These should: – Be practically oriented ("hands-on"). – Involve collaboration and/or learner communities. – Align with students' needs. – Model, or scaffold, effective assessment practices. Increase student involvement in formative assessment and, more specifically, improve students' proficiency in working with formative assessment. This can be done through a combination of actions including: – Ensuring students provide each other with feedback at appropriate times. – Creating safe environments. – Using technology to make assessment tasks more manageable.

and PE grades), while in Modell and Gerdin's (2022) (26) research, some students believed that it was not specifically gender but certain physical characteristics such as strength and speed that were advantaged in PE. In other investigations, students claimed that assessment was unfair because they were assessed on skills and knowledge they had not had the opportunity to learn during lessons (26, 27, 32). One large-scale statistical analysis of grades supports this contention, indicating that prior sporting experience is likely to result in higher grades (38). Noteworthy with respect to equity is that research suggests that other factors such as ethnicity and parents' educational level (38) and relative age (33) also influence grades, yet students have *not* associated these factors with assessment inequity.

Another less commonly identified subproblem concerning students' negative experiences of assessment involved stress (13). Students in Røset et al.'s (2023) (34) investigation claimed that the highly visible ways in which assessment tasks were conducted in Norwegian PE lessons led to anxiety and worry, and feelings of being surveilled. Kim and Lee (2022) (13) observed that, despite experiencing highly stressful assessment practices as school students, PE teachers were inclined to engage in those same practices when teaching.

Significant with respect to students' negative experiences is that few studies have indicated that students cite lack of transparency as a problem (26 constitutes an exception). This is noteworthy in light of the abundance of research problematizing teachers' personalized assessment practices (11, 24, 33, 34, 47). One investigation even indicated that students feel relatively confident in recognizing the criteria on which they are being assessed (32). This confidence may be misplaced, although it is also possible that students recognize and react to their teachers' use of personal, "unofficial" criteria for assessment.

The third problematic area concerns direction provided by state curricula in matters of assessment. On one hand, Spanish researchers have claimed that curricula are becoming closed and prescriptive, and the State is emphasizing standardization and measurement (29). On the other hand, (mainly Scandinavian) researchers have contended that educational policy is too open and has provided teachers with insufficient guidance and direction (15, 32). Svennberg (2017) (37) claimed that national curricula do not provide clear distinctions between grade descriptions and Svennberg et al. (2014: 211) (39) asserted that "as long as the grading criteria do not address classroom realities and restricting factors, the grades will continue to be inconsistent with the [curricular] criteria." In an investigation that provides some insights into why educational policy might fall short of teachers' expectations, Scanlon et al. (2019) (35) observed how different stakeholders influenced the construction of the Irish curriculum. The researchers concluded that teachers had markedly different expectations to policy makers and that tensions resulted in compromises for all parties.

Potential solutions to problems with summative assessment

Although researchers have critiqued many aspects of summative assessment, the logic underpinning solutions is remarkably consistent across the corpus of literature. The principal solution is to provide teachers with training either through (better) initial teacher education (7, 25) or, more commonly, through continuing professional development (CPD—24, 31). More precisely, researchers have suggested that teachers need to learn to be more reflective so they can "see," and thus prevent, their perceptions from "contaminating" (11: 161) their assessment of student learning (39).³ Teachers need to develop a sound understanding of curricular intentions so they are familiar with what students are supposed to learn (24, 32). Once they understand curricular intentions, they need to ensure that their teaching and assessment practices are aligned with those intentions (7, 46). Redelius and Hay (2009) (31) suggest that to ensure this alignment, teachers must

learn to write assessment statements, to collect evidence, and to understand and use this evidence. In the classroom, teachers must be able to communicate learning intentions and assessment criteria clearly to students (11, 27, 37). Zhu (2015: 419) (48) suggests, for example, that PE teachers “should explicitly teach the achievement standards and grade expectations [to students].” While scholars have placed considerable confidence in CPD to improve assessment practices (11, 31), the responsibility of education agencies to provide teachers with clear guidelines that explain how assessment should be carried out has also been cited (11, 32).

Other potential solutions offer alternative ways of understanding the antecedents of problems associated with summative assessment. Marmeleira et al. (2020) (23), for example, propose that boys receive higher grades than girls in PE because girls have lower levels of sport participation outside of school and are thus less familiar with the content dealt with in PE. Their solutions revolve around making PE more accommodating for girls by introducing, for example, single-sex competition and noncompetitive environments. In their discussion of tensions around curricular development for senior PE, Scanlon et al. (2019) (35) note that there is a need for stakeholders to work democratically to determine what should be assessed. And in response to the claim that teachers’ personal criteria for “good performance” contaminate the assessment process, Svennberg et al. (2018) (40) offer an alternative to CPD when they contend that it is impossible to separate behavioral aspects such as attendance and being on-task from knowledge and skills. More specifically, they suggest that it is not possible to comment on students’ development of knowledge if they are not in class or are not attempting to demonstrate the knowledge being learned during lessons. In this respect, they recommend accepting a relationship between behavioral aspects and learning.

Problems with formative assessment

Problems identified with formative assessment can also be divided into three broad “areas,” each of which contains a collection of subproblems. These areas revolve around: (1) teachers’ lack of competence with formative assessment practices; (2) challenges that teachers experience with formative assessment; and (3) problems that students experience with formative assessment.

Scholars have identified many shortcomings with teachers’ formative assessment practices, often referring to teachers’ lack of assessment literacy (9, 17). Slingerland et al. (2024) (36) claimed that teachers lacked proficiency in developing learning targets or success criteria, rarely provided students with information on how to improve their performance and were sometimes unsure of what assessment strategy they were using. Leirhaug and Annerstedt (2016) (15) proposed that the teachers in their investigation constantly provided students with feedback, but that the comments did not relate to specific learning goals, while Chng and Lund (2019) (8) suggested that the teacher in their study failed to place much emphasis on peer feedback.⁴ Tolgfors (2018) (41) pointed to considerable variation between teachers, suggesting that assessment for learning took different forms in PE, and that often the learning motive for assessment was overpowered by the accountability motive (essentially turning assessment intended for learning into assessment for grade generation—see also 44). Raising the question of teachers’ ability to align different forms of assessment, Leirhaug (2016) (14) observed an apparent paradox, asserting that better performance in formative assessment tasks was not associated with better performance in summative assessment tasks. One study found that a quarter of the 210 PE teachers surveyed claimed not to use formative assessment strategies at all (25).

Given reports of teachers’ lack of assessment knowledge, it is perhaps unsurprising that teachers have expressed apprehension about conducting formative assessment. Some teachers’ concerns are reflected in their reluctance to deviate from traditional PE practices (9). Teachers are worried that

more time dedicated to formative assessment will result in less time for physical activity (36). Indeed, teachers in many studies expressed uneasiness at the amount of time taken up by formative assessment (4, 9, 10, 19, 22, 45). In addition to concerns about time, teachers in Slingerland et al.'s (2024) (36) investigation were worried that the increased student autonomy involved in the formative assessment tasks offered them less control and could raise safety issues when done in the context of gymnastics (see also 21). Other studies indicated that teachers lacked confidence either in their own assessment capacities (9) and/or in their students' capacities to engage in assessment practices (36). Tolgfors et al. (2022) (43) noted that even if teachers have developed a sound understanding of formative assessment principles in initial teacher education and intend to practice formative assessment once they are in schools, they may find it difficult to implement when faced with oppositional expectations from colleagues and students.

A third problematic area relates to students and their lack of experience with formative assessment. Some students in Leirhaug and Annerstedt's (2016) (15) research were unsure of whether they had engaged in formative assessment and claimed to have minimal experience with self- or peer-assessment in PE (see also 20). Students in Butler and Hodge's (2001) (6) investigation reported that they received general feedback from their peers but not specific feedback when working with assessment for learning. In a similar vein, Chng and Lund (2019) (8) observed pupils rushing to complete assessments so they could continue playing games or leave PE for the next class, while Slingerland et al. (2024) (36) noted that the students they observed were simply not particularly good at giving or receiving feedback. Regarding feedback, Backman et al. (2024: 11) (3) identified a slightly different problem, pointing out that students' physical capital affected "who is allowed to give feedback to whom." Their results suggest that students with higher status have more to say in peer assessment situations. In this respect, Backman et al.'s (2024) (3) research suggests that, even if students are familiar with technical aspects of formative assessment, interpersonal issues still need to be managed (see also 15 for reference to student marginalization).

Potential solutions to problems with formative assessment

Unlike the solutions offered to improve summative assessment, scholars' recommendations for improving formative assessment are diverse. Some scholars have adopted a philosophical position, suggesting that tensions will always make formative assessment complex. Tolgfors (2019) (42), for example, draws attention to the influence of sociocultural factors, claiming that assessment practices need to be adapted to the context in which they are taking place. In a separate paper, Tolgfors (2018) (41) suggested that teachers need to employ different versions of formative assessment and that there is no standardized way of doing formative assessment that will be effective in all situations (see also 28, 43 and 44).

Other scholars have assumed more practically oriented positions. As well as broad calls for greater student involvement (4, 16, 22, 42), researchers have made an array of concrete suggestions that are supposed to help students work with formative assessment (1, 8). Lorente-Catalán and Kirk (2016) (18), for instance, suggest that teachers should introduce peer-assessment before self-assessment, reasoning that self-assessment is more difficult and engaging in peer-assessment initially will give students confidence. The same researchers suggest that teachers should give feedback prior to awarding grades in order to place greater emphasis on learning and reduce the emphasis on grades. They further propose that students should be provided with assessment criteria when they are doing self-assessment (18). Fletcher et al. (2024) (9) and Butler and Hodge (2001) (6) provide similar practical recommendations for teachers that include creating safe and trusting environments, using technology to make assessment less time consuming and easier to manage, and

maintaining tight control of situations where students are assessing one another. Barrientos Hernán et al. (2023) (4) reported on the practical solutions developed by the teachers in their investigation. These included recording information on a few students in every lesson and using time spent moving between classrooms to hold personal assessment interviews with students and thus protect active learning time in lessons.

As with summative assessment, scholars have underscored the need for teachers to learn more about formative assessment through CPD (4, 18). Slingerland et al. (2024) (36) claimed that, while teachers may resolve some problems in the course of using formative assessment, professional development has the potential to challenge teachers' fundamental understandings of assessment and provide important opportunities to consider how formative assessment can work in their particular contexts. They outline many characteristics of effective CPD, suggesting that it should involve a "hands-on" approach where teachers can work collaboratively and/or in communities, and focus on content that is aligned with students' needs. Supporting Slingerland et al.'s (2024) (36) claim, Ní Chróinín and Cosgrave (2013) (28) reported that attending postgraduate studies in PE radically altered primary school teachers' understandings and practices of formative assessment. Ní Chróinín and Cosgrave (2013) emphasized the value of scaffolding effective assessment strategies, concluding that initial teacher education needs to place more emphasis on assessment and sample assessment tools, and that teachers then need to be supported through CPD.⁵

Discussion: the persistence of problems with summative and formative assessment

Findings from the review suggest that the problems surrounding summative and formative assessment are persistent and that an assessment culture is still lacking in PE. Studies show that teachers continue to account for learning using factors such as participation and effort (34, 37) rather than official learning intentions, a pattern that extends back to the 1980s and 1990s (Matanin and Tannehill, 1994; Veal, 1988). Such assessment practices continue to raise concerns around instructional alignment, transparency and comparability (AIESEP, 2020). Studies also show that students continue to experience summative assessment and specifically grading as inequitable (26, 27, 32), and that teachers express concerns relating to classroom management and physical activity levels when enacting formative assessment (6, 8, 15, 36).

There are at least four possible reasons why assessment research has not helped physical educators move on from these problems. First, despite questions raised with respect to educational policy (29, 37, 39), scholarship has assigned the bulk of the responsibility to teachers. With regard to both formative and summative assessment, scholars have presented teachers as either unable or unwilling to conduct assessments effectively. With this framing, teacher educators and providers of CPD have assumed prime importance, the assumption being that if these actors can help teachers to learn about assessment, teachers will develop the competence and desire to assess effectively in the classroom (4, 18). Yet researchers' confidence in physical education teacher education (PETE) and CPD is based on the assumption that PETE lecturers and CPD providers are assessment experts, an assumption that warrants investigation. Furthermore, relatively few studies provide detailed descriptions of what CPD should look like, especially concerning summative assessment. Confidence in PETE and CPD also appears somewhat optimistic in light of a long line of teacher socialization studies suggesting that PETE struggles to make the impact on beginning teachers' practices that PETE educators hope for (43; Starck et al., 2018; see also Schempp et al., 1993; Templin et al., 2016), and that in most national contexts, teachers receive modest opportunities

for CPD (Tannehill et al., 2021).⁶ Indeed, some PE scholars have suggested that, even when teachers have developed assessment literacy during their initial teacher education they still find it difficult to use when they enter school cultures in which certain forms of assessment have been either absent or are actively resisted by colleagues and students (43). Additionally, there are few guarantees that CPD provision will reach the teachers that need it (24). In short, despite some studies demonstrating the considerable positive impact that both PETE and CPD can have on teachers' assessment practices (18, 28, 36), there are challenges that need to be addressed if they are to be successful.

A second reason that helps to explain assessment scholarship's limited success with helping teachers resolve longstanding problems is contextual diversity. Even if the problems identified in the synthesis occur across national contexts, the conditions that constrain and afford assessment practices in those contexts vary considerably (Tolgfors and Barker, 2023). For example, while senior PE is examinable in many Anglophone and Scandinavian countries, and while final grades in the subject will help to determine students' possibilities for tertiary education and may thus be considered "high stakes" (Hay and Penney, 2013), students encounter PE in very different conditions. Senior students in countries following the British model will take PE as one of a complement of three to five school subjects, will find themselves with peers who have specifically chosen to study PE, and will most likely be assessed with some form of external examination (30, 35). In Scandinavian schools, where education has a broad focus even in senior years, senior students will take PE as one of between 15 and 20 school subjects—all of which will affect university entry—will participate in PE in heterogenous groups with peers who have varying levels of interest in the school subject, and will not experience any external examination in PE (14–17, 26, 27). Australian and Norwegian PE teachers will, for example, spend widely varying amounts of time with their students and have markedly different responsibilities regarding both formative and summative assessment. The upshot is that the solutions offered by Scandinavian PE scholars for common assessment problems may have limited transferability or even relevance to Australian or British physical educators, for example, and vice versa.

A third reason that helps to explain assessment scholarship's limited success is the disconnection between formative and summative assessment often present in the literature. In 42 of the 48 studies reviewed, researchers treated formative and summative assessment activities as discrete, implicitly suggesting that one can be manipulated without impacting the other (see 5, 10, 20, 25, 29, 46 for exceptions). This understanding is reflected in the fact that summative assessment is critiqued mainly for the intrusion of teachers' personal and internalized performance criteria while formative assessment's "main problem" concerns teachers' lack of proficiency. To some extent, differentiating between the two types of assessment conceptually and practically is justified: they have different purposes and generally occur in different stages of learning (Otero-Saborido et al., 2021). At the same time, the heavy emphasis placed on instructional alignment (26, 27) calls a clear separation into question (see AIESEP, 2020). If instructional alignment entails students first learning and then being held accountable for that learning, then effective formative assessment strategies must be seen as an integral part of summative assessment (14, 15). We should emphasize that we are not suggesting that instructional alignment necessarily *should* characterize pedagogical systems. We have proposed elsewhere that aligning curriculum, pedagogy and assessment too rigidly may close down potential learning trajectories (Tolgfors and Barker, 2023). We are suggesting more broadly that researchers need to provide ways to fit summative and formative assessment practices together if change is to take place.

A fourth and final factor that may explain why problems with assessment persist despite the burgeoning amount of assessment research concerns how assessment problems relate to one another.

Practically all of the problems presented are not isolated issues but are rather “entangled” with one another. In the case of summative assessment, teachers’ use of personal and internalized criteria, and lack of instructional alignment are both the result of, and a contributor to, a lack of transparency. Furthermore, if curricula offered more direction on assessment matters or conversely, if teachers better understood curricular intentions, there would be less room for teachers to use personal criteria in the first place. Indeed “entanglement” is an apposite term because it draws attention to the need to consider multiple problems simultaneously. Entanglement also invites us to examine where “knots” reside. The idea that curricular statements might be either too narrow and prescriptive (29), or too vague and ambiguous (32, 39), constitutes one such knot. On one hand, precision and detail in curricular documentation are presented as positive and worthy to strive toward, contributing to transparency, equity and comparability. On the other, precision and detail are presented as prescriptive and a barrier to teacher autonomy. The conflict is thus between two competing, and in our view, compelling, ways of understanding the role of curricula in guiding assessment practices. If policy makers would resolve one problem entirely, by having say, a detailed, comprehensive prescription for assessment (29) that was mandatory for teachers to follow, teachers could be left with little room for exercising professional judgment. The question is thus not “how do we achieve problem-free assessment?” but to paraphrase educational philosopher Joseph Dunne, “how do we engage in assessment in ways that those involved will find it satisfactory, or at least not entirely unacceptable?” (Dunne, 2005: 381).

Conclusion

We do not expect researchers to provide silver bullets for practical problems. We understand that the relationship between research findings and classroom practices is nonlinear at best. We would also like to point out that synthesis reviews such as this provide a wide-lens picture of research conducted internationally. In synthesizing a collection of studies, such reviews also *conceal* problems that are embedded in specific national or regional contexts. At the same time, we are concerned that after sustained research over more than two decades, PE researchers in the mid-2020s are still identifying similar types of problems and offering similar strategies for improvement to their counterparts at the turn of the millennium.

Based on our synthesis, we see several ways forward. First, we believe that PE scholars need to continue to problematize assessment in different ways. New ways of locating assessment challenges, especially ways that go beyond individualist understandings based on lack of competence or willingness, will lead to alternative research questions, methodologies, and proposed solutions. Scholars need to investigate, for example, how certain content becomes difficult to assess (see Hay and Penney, 2009; Thorburn, 2007, for conceptual discussions of this topic), how certain types of assessment knowledge are taught in PETE and CPD, and how assessment policy changes over time. In addition to providing a basis for accounts of student learning, summative assessment is supposed to generate data that can be used for the evaluation and development of PE programs, yet we found few investigations into this role of assessment in our synthesis, let alone problematizations. These are all topics that have the potential to extend physical educators’ understandings of assessment.

Second, although we have raised questions about CPD’s capacity to resolve problems, very little research has been conducted in this area. What appears especially important is that CPD is not presented as an afterthought in investigations of assessment, or as a hypothetical strategy to solve problems in the future, but instead as a starting point for investigations (see, for e.g., 28, 36). Research into CPD—and PETE for that matter—necessitates going beyond descriptive methodologies, and

including intervention- and intervention-type approaches that can help researchers to investigate processes and change over time.⁷ If researchers do propose CPD as a possible solution to problems that they have identified, then they should at least be explicit on how they think it should be done.

Finally, with reference to the tension between curricular clarity and teacher autonomy (see 15, 29, 32, 35, 37), we suggested that if physical educators resolve some problems, they may well introduce others. This is not to suggest that physical educators should simply concede that no way of assessing is perfect or that anything goes. On the contrary, it is to emphasize that physical educators need a particularly thorough understanding of how different political and philosophical perspectives pull assessment practices in different directions (Penney et al., 2009; Rink and Mitchell, 2002; Starck et al., 2018), of the advantages and disadvantages of the assessment strategies they have available (MacPhail et al., 2023; Melograno, 1998; Scanlon et al., 2023), of the conditions that facilitate the successful implementation of assessment (Hay and Penney, 2009), and importantly, of the content with which they are dealing (Ward et al., 2015). Such holistic understandings—captured in the notion of assessment literacy (9, 17, DinanThompson and Penney, 2015; Starck et al., 2018)—appear key if physical educators are to change assessment practices.

We want to finish with several brief reflections regarding the limitations of our review. We have concentrated on recurring problems. Scholars cited an array of problems but if they showed up only a handful of times, they have not been discussed here.⁸ In one respect, our review consequently lacks nuance. In another, many infrequently mentioned problems are permutations of the problems we have presented. Our sense is that we have not missed problems that are completely unrelated to the ones we have shown. Second, we have included little information about the content that was being assessed in the investigations or the kinds of pedagogies that were being used. This kind of detail could certainly have added valuable context from which to interpret the findings and is something that should be kept in mind in future reviews. Finally, our synthesis has been strongly shaped by our focus on problems and the solutions offered for those problems. As a result, we have not developed a detailed picture of the advances, developments, and improvements that are taking place in PE (see 28). There is consequently a need for a comprehensive consideration of the progress that has been made with respect to assessment in PE.

ORCID iDs

Dean Barker  <https://orcid.org/0000-0003-4162-9844>

Björn Tolgfors  <https://orcid.org/0000-0002-1773-7792>

Annica Caldeborg  <https://orcid.org/0000-0003-1097-8750>

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was funded by Centrum för Idrottsforskning (ID: P2025-0095).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

Notes

1. Interestingly, Rink and Mitchell (2002: 210) also noted the risk of teachers “teaching to the test” but claimed that even teaching to the test could constitute a “substantive improvement for many programs.”
2. School levels are organized in different ways in different countries, and local terminology is used to refer to levels throughout the assessment literature. In categorizing the articles, we grouped together studies that focused on “primary,” “elementary,” “grades 1–6,” and “students aged 12 or younger.” We also grouped together “secondary,” “upper secondary,” “junior high school,” “lower high school,” “upper high school,” “grades 7–13” and “students aged 13 or older.” We recognize that there are differences between forms of PE conducted in “lower high school” and “upper high school” (especially with respect to forms of examination); however, since these differences are not universal, it was deemed inappropriate to divide “high school” into two separate categories. This aspect is discussed in the Discussion.
3. Penney et al. (2023) (30) describe an alternative solution to the problem of teachers’ use of personal criteria whereby the same external evaluators travelled around schools in Western Australia to conduct assessment in the same way. The scholars noted, nonetheless, that assessment conditions were still not entirely comparable since the type and quality of equipment, and the size and composition of student groups at the schools varied.
4. Importantly, the teacher in this study did use the data from the formative assessments for lesson planning. The students recognized that he was looking at the results of these assessments, providing a measure of accountability (8).
5. One anonymous reviewer commented here that initial teacher education needs to have a programmatic emphasis on using assessments and that preservice teachers need to be held accountable for learning about assessment.
6. Tannehill et al. (2021) surveyed official requirements for CPD in 14 European countries. The median number of hours provided by the 10 countries that offered CPD was 24 per year.
7. Only six of the investigations in the synthesis relied explicitly on an intervention methodology (8, 12, 21, 28, 36, 45). In three investigations, scholars attempted to introduce and monitor innovation but did not explicitly refer to an “intervention” (3, 9, 18). Despite using a range of methods for data production, the rest of the studies could broadly be described as “descriptive.”
8. Redelius and Hay (2009) (31), for example, noted that summative assessment tended to shape students’ understanding of themselves, sometimes in negative ways. Røset et al. (2023) (34) observed that increased focus on summative assessment appeared to reduce students’ enjoyment.

References

- AIIESEP. (2020) *The AIIESEP position statement on physical education assessment*. Available at: <http://aiesep.org/scientific-meetings/position-statements/> (accessed 20 August 2024).
- Backman E, Quennerstedt M, Tolgfors B, et al. (2024) Peer assessment in physical education teacher education—A complex process making social and physical capital visible. *Curriculum Studies in Health and Physical Education* 15(3): 274–288.
- Barrientos Hernán EJ, López-Pastor VM, Lorente-Catalán E, et al. (2023) Challenges with using formative and authentic assessment in physical education teaching from experienced teachers’ perspectives. *Curriculum Studies in Health and Physical Education* 14(2): 109–126.
- Black P and Wiliam D (2009) Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability* 21(1): 5–31.
- Bores-García D, Hortigüela-Alcalá D, González-Calvo G, et al. (2020) Peer assessment in physical education: A systematic review of the last five years. *Sustainability* 12(21): 9233.
- Bullock HL, Lavis JN, Wilson MG, et al. (2021) Understanding the implementation of evidence-informed policies and practices from a policy perspective: A critical interpretive synthesis. *Implementation Science* 16: 1–24.

- Butler SA and Hodge SR (2001) Enhancing student trust through peer assessment in physical education. *The Physical Educator* 58(1): 30–39.
- Chng L and Lund J (2019) The impact of formative assessment on the success and response rate in a 6th grade physical education badminton unit. *International Journal of Kinesiology in Higher Education* 3(1): 12–22.
- Chng LS and Lund J (2018) Assessment for learning in physical education: The what, why and how. *Journal of Physical Education, Recreation & Dance* 89(8): 29–34.
- Collier D (2011) Increasing the value of physical education: The role of assessment. *Journal of Physical Education, Recreation & Dance* 82(7): 38–41.
- Depraetere J, Vandeviver C, Keygnaert I, et al. (2021) The critical interpretive synthesis: An assessment of reporting practices. *International Journal of Social Research Methodology* 24(6): 669–689.
- DinanThompson M and Penney D (2015) Assessment literacy in primary physical education. *European Physical Education Review* 21(4): 485–503.
- Dixon-Woods M, Cavers D, Agarwal S, et al. (2006) Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology* 6: 1–13.
- Dunne J (2005) An intricate fabric: Understanding the rationality of practice. *Pedagogy, Culture & Society* 13(3): 367–390.
- Farias L and Laliberte Rudman D (2016) A critical interpretive synthesis of the uptake of critical perspectives in occupational science. *Journal of Occupational Science* 23(1): 33–50.
- Fletcher C, Iannucci C and Scanlon D (2024) A teacher's self-study of digitally-enabled assessment practices to support enhancements in assessment literacy in primary physical education. *Curriculum Studies in Health and Physical Education* 15(3): 255–273.
- Grehaigue J-F, Godbout P and Bouthier D (1997) Performance assessment in team sports. *Journal of Teaching in Physical Education* 16(4): 500–516.
- Hay PJ and Penney D (2009) Proposing conditions for assessment efficacy in physical education. *European Physical Education Review* 15(3): 389–405.
- Hay PJ and Penney D (2013) *Assessment in Physical Education: A Sociocultural Perspective*. London: Routledge.
- Herrero-González D, López-Pastor VM, Manrique-Arribas JC, et al. (2024) Formative and shared assessment: Literature review on the main contributions in physical education and physical education teacher education. *European Physical Education Review* 30(3): 493–510.
- Jean de Dieu H and Zhou K (2021) Physical literacy assessment tools: A systematic literature review for why, what, who, and how. *International Journal of Environmental Research and Public Health* 18(15): 7954.
- Kim Y and Lee O (2022) Exploring the phenomenological meanings of assessment in physical education through primary teachers' lived experiences. *Physical Education and Sport Pedagogy* 27(6): 578–590.
- Lander NJ, Barnett LM, Brown H, et al. (2015) Physical education teacher training in fundamental movement skills makes a difference to instruction and assessment practices. *Journal of Teaching in Physical Education* 34(3): 548–556.
- Leirhaug PE (2016) Exploring the relationship between student grades and assessment for learning in Norwegian physical education. *European Physical Education Review* 22(3): 298–314.
- Leirhaug PE and Annerstedt C (2016) Assessing with new eyes? Assessment for learning in Norwegian physical education. *Physical Education and Sport Pedagogy* 21(6): 616–631.
- López-Pastor VM, Kirk D, Lorente-Catalán E, et al. (2013) Alternative assessment in physical education: A review of international literature. *Sport, Education and Society* 18(1): 57–76.
- Lorente-Catalán E and Kirk D (2016) Student teachers' understanding and application of assessment for learning during a physical education teacher education course. *European Physical Education Review* 22(1): 65–81.

- MacPhail A, Tannehill D, Leirhaug PE, et al. (2023) Promoting instructional alignment in physical education teacher education. *Physical Education and Sport Pedagogy* 28(2): 153–164.
- Marmeleira J, Folgado H, Martinez Guardado I, et al. (2020) Grading in Portuguese secondary school physical education: Assessment parameters, gender differences and associations with academic achievement. *Physical Education and Sport Pedagogy* 25(2): 119–136.
- Matanin M and Tannehill D (1994) Assessment and grading in physical education. *Journal of Teaching in Physical Education* 13(4): 395–405.
- Melograno V (1998) *Professional and Student Portfolios for Physical Education*. Champaign: Human Kinetics.
- Modell N and Gerdin G (2022) ‘But in PEH it still feels extra unfair’: Students’ experiences of equitable assessment and grading practices in physical education and health (PEH). *Sport, Education and Society* 27(9): 1047–1060.
- Moura A, Graça A, MacPhail A, et al. (2021) Aligning the principles of assessment for learning to learning in physical education: A review of literature. *Physical Education and Sport Pedagogy* 26(4): 388–401.
- Ní Chróinín D and Cosgrave C (2013) Implementing formative assessment in primary physical education: Teacher perspectives and experiences. *Physical Education and Sport Pedagogy* 18(2): 219–233.
- Oslin JL, Mitchell SA and Griffin LL (1998) The game performance assessment instrument (GPAI): Development and preliminary validation. *Journal of Teaching in Physical Education* 17(2): 231–243.
- O’Sullivan M (2013) New directions, new questions: Relationships between curriculum, pedagogy, and assessment in physical education. *Sport, Education and Society* 18(1): 1–5.
- Otero-Saborido FM, Torreblanca-Martínez V and González-Jurado JA (2021) Systematic review of self-assessment in physical education. *International Journal of Environmental Research and Public Health* 18(2): 766.
- Penney D, Brooker R, Hay P, et al. (2009) Curriculum, pedagogy and assessment: Three message systems of schooling and dimensions of quality physical education. *Sport, Education and Society* 14(4): 421–442.
- Penney D, O’Hara E and Lund R (2023) Enhancing quality and equity? Performance assessment validation in examination physical education in Western Australia. *Curriculum Studies in Health and Physical Education* 14(3): 288–305.
- Redelius K and Hay P (2009) Defining, acquiring and transacting cultural capital through assessment in physical education. *European Physical Education Review* 15(3): 275–294.
- Rink J and Mitchell M (2002) High stakes assessment: A journey into unknown territory. *Quest (Grand Rapids, Mich)* 54(3): 205–223.
- Røset L, Green K, Sigurjonsson T, et al. (2023) Now we have gym, now we have to perform’: Norwegian students’ perceptions of assessment and grading in physical education. *European Educational Research Journal* 23(6): 879–901.
- Scanlon D, MacPhail A and Calderón A (2019) Original intentions and unintended consequences: The ‘contentious’ role of assessment in the development of leaving certificate physical education in Ireland. *Curriculum Studies in Health and Physical Education* 10(1): 71–90.
- Scanlon D, MacPhail A, Walsh C, et al. (2023) Embedding assessment in learning experiences: Enacting the principles of instructional alignment in physical education teacher education. *Curriculum Studies in Health and Physical Education* 14(1): 3–20.
- Schempp PG, Sparkes AC and Templin TJ (1993) The micropolitics of teacher induction. *American Educational Research Journal* 30: 447–472.
- Siedentop D and Tannehill D (2000) *Developing Teaching Skills in Physical Education*. Mountain View, CA: Mayfield.

- Slingerland M, Weeldenburg G and Borghouts L (2024) Formative assessment in physical education: Teachers' experiences when designing and implementing formative assessment activities. *European Physical Education Review* 30(4): 620–637.
- Starck JR, Richards KAR and O'Neil K (2018) A conceptual framework for assessment literacy: Opportunities for physical education teacher education. *Quest (Grand Rapids, Mich)* 70(4): 519–535.
- Sundaresan N, Dashoush N and Shangraw R (2017) Now that we're "well rounded," let's commit to quality physical education assessment. *Journal of Physical Education, Recreation & Dance* 88(8): 35–38.
- Svennberg L (2017) Swedish PE teachers' understandings of legitimate movement in a criterion-referenced grading system. *Physical Education and Sport Pedagogy* 22(3): 257–269.
- Svennberg L, Meckbach J and Redelius K (2014) Exploring PE teachers' 'gut feelings': An attempt to verbalise and discuss teachers' internalised grading criteria. *European Physical Education Review* 20(2): 199–214.
- Svennberg L, Meckbach J and Redelius K (2018) Swedish PE teachers struggle with assessment in a criterion-referenced grading system. *Sport, Education and Society* 23(4): 381–393.
- Tannehill D, Demirhan G, Čaplová P, et al. (2021) Continuing professional development for physical education teachers in Europe. *European Physical Education Review* 27(1): 150–167.
- Templin TJ, Padaruth S and Sparkes AC (2016) A historical overview of teacher socialization in physical education. In: Richards K and Gaudreault K (eds) *Teacher Socialization in Physical Education*. London: Routledge, pp. 11–31.
- Thorburn M (2007) Achieving conceptual and curriculum coherence in high-stakes school examinations in physical education. *Physical Education and Sport Pedagogy* 12(2): 163–184.
- Tolgfors B (2018) Different versions of assessment for learning in the subject of physical education. *Physical Education and Sport Pedagogy* 23(3): 311–327.
- Tolgfors B (2019) Transformative assessment in physical education. *European Physical Education Review* 25(4): 1211–1225.
- Tolgfors B and Barker D (2023) The glocalization of physical education assessment discourse. *Sport, Education and Society* 28(1): 1–16.
- Tolgfors B, Quennerstedt M, Backman E, et al. (2022) Enacting assessment for learning in the induction phase of physical education teaching. *European Physical Education Review* 28(2): 534–551.
- Veal ML (1988) Pupil assessment perceptions and practices of secondary teachers. *Journal of Teaching in Physical Education* 7(4): 327–342.
- Ward P, Kim I, Ko B, et al. (2015) Effects of improving teachers' content knowledge on teaching and student learning in physical education. *Research Quarterly for Exercise and Sport* 86(2): 130–139.
- Zhu X (2015) Student perspectives of grading in physical education. *European Physical Education Review* 21(4): 409–420.

Author biographies

Dean Barker holds a position as Associate Professor in Physical Education at the School of Health Sciences at Örebro University, Sweden. His research focuses on pedagogical strategies that help PE teachers enhance students' learning experiences.

Björn Tolgfors holds a position as Associate Professor in Physical Education at the School of Health Sciences at Örebro University, Sweden. His research focuses on teaching, learning and assessment in school PE and PETE.

Annica Caldeborg holds a position as Lecturer in Physical Education at the School of Health Sciences at Örebro University, Sweden. Her research explores teaching, learning, and students' perspectives within PE and health.

Appendix

Appraisal prompts for informing judgments about quality of papers (Dixon-Woods, 2006)

Are the aims and objectives of the research clearly stated?

Is the research design clearly specified and appropriate for the aims and objectives of the research?

Do the researchers provide a clear account of the process by which their findings were produced?

Do the researchers display enough data to support their interpretations and conclusions?

Is the method of analysis appropriate and adequately explicated?