Ed Process Int J

www.edupij.com

ISSN 2147-0901 (Print), 2564-8020 (Online)

EDUCATIONAL PROCESS
INTERNATIONAL JOURNAL

Peer Observation of Teaching in Higher Education: Systematic Review of Observation Tools

Fernando Manuel Otero Saborido, José Antonio Domínguez-Montes, José Manuel Cenizo Benjumea, Gustavo González-Calvo

Research Article

Cite this article: Otero-Saborido, F.M., Domínguez-Montes, J.A., Cenizo-Benjumea, J.M., & González-Calvo, G. (2024). Peer Observation of Teaching in Higher Education: Systematic Review of Observation Tools. *Educational Process: International Journal*, 13(1): 84-101. https://doi.org/10.22521/edupij.2024.131.6

Received September 14, 2023
Accepted January 04, 2024
Published Online February 29, 2024

Keywords:
Instruments, peer partner, education tertiary, mentoring

Author for correspondence:
José Manuel Cenizo-Benjumea
jmcenben@upo.es
Teaching of physical education, Pablo de Olavide University, Sevilla, España.

Abstract

Background/purpose. This study presents a systematic review of teaching observation instruments in the current literature based on PRISMA standards.

Materials/methods. Three researchers performed searches on two databases, SCOPUS and Web of Science, focusing on two criteria: a) peer observation of teaching and b) higher education, with search terms included in the "Title/Keyword" fields. The AND command was used to join certain words, including peer observation and teaching, whilst the OR command was used to separate search terms within each criterion. Five exclusion criteria were defined and applied following the initial searches. The quality of research conducted in the literature using observation tools was assessed using a validated instrument in social science research.

Results. The results revealed a total of 13 instruments that were analyzed in terms of four variables: country, validation, observation, and feedback. a) Country: More than half were designed by researchers from universities in the United States and Australia. b) Validation: Only three studies were designed following some kind of validation procedure. c/d) Observation and feedback: The number of items ranged from very loosely structured, with only a few items, to more comprehensive research. The most repeated item (8 of 13 instruments) was about the objectives of the observation section. Four study instruments included only an observation section, with no specific feedback section. Of the remainder, some included all three aspects of "strengths," "weaknesses," and "comments" in the feedback section, while others included only a feedback section.

Conclusion. Excessive question numbers could make observation exercises overly complex, unless the items are distributed and observed across several sessions. An appropriate number of questions would correspond to the amount deemed by teachers themselves to be essential to observe the teaching process. Observation tools should include fields in which observers may add qualitative comments to deepen the understanding of the record and to improve the feedback quality.

1. Introduction

A substantial body of scientific evidence exists on the positive correlation between feedback and learning, as supported in the research published by Brooks, Carroll, et al. (2019), Brooks, Huang, et al. (2019), Hattie and Clarke (2018), and Panadero and Lipnevich (2022). The current study, however, focuses specifically on feedback provided through peer observation of teaching (POT), where teacher A observes teacher B while teacher B is teaching, and subsequently provides feedback to teacher B on their teaching performance in the classroom. Although no studies were found in the literature exclusively on POT applied in the higher education context, early evidence quantifying the impact of POT on learning exists for other educational stages (Burgess et al., 2021). POT has become increasingly popular in higher education institutions worldwide, including in the United States, Australia, and the United Kingdom (Carragher & McGaughey, 2016; Johnston et al., 2022).

It would be inappropriate to analyze POT without considering the rationales that underlie the conceptual POT frameworks, which are: technical, practical, and critical. To synthesize, technical rationality is behaviorist and quantitative in nature, with a curricular approach formed according to pedagogically-based objectives, whilst practical rationality refers to a process-based and student-centered pedagogy. Finally, critical rationality is based on the philosophy of practical rationality, but is oriented more towards transforming reality (López Pastor, 1999). Analysis of research by different authors shows how different interpretations of POT align with these three rationales (Bell & Mladenovic, 2008; Byrne et al., 2010; Gosling, 2002; Peel, 2005).

Bell and Mladenovic's (2008) theoretical framework rests on two continuums: "performance/development and training purposes on the vertical axis" and "formal-informal processes on the horizontal axis," which were originally suggested by Peel (2005). Peel (2005) also proposed a conceptual structure that organized POT into three dimensions. The first dimension (D1) is based on a technical rationality that links teachers from an informative viewpoint, whilst the second dimension (D2) relates teachers based on collaboration and process-based research. The last dimension (D3) focuses on critical reflection as well as moral and ethical criteria. The literature shows that Gosling (2002) and Byrne et al. (2010) used different but related models. Gosling (2002) created three POT models: the evaluation model, the professional development model, and the collaborative model. The evaluation model aims to produce judgments on teaching practices, while the professional development and the collaborative models were designed to improve didactic skills. In contrast, Byrne et al. (2010) outlined four models. The first model identifies underperformance and is based on authority, while the collaborative models seek engagement and discussion about practice. Byrne et al. (2010), however, added a fourth model that incorporated "ideas about learning and teaching into practice through a shared and reciprocal process that has potential for greater impact than can be gained via a one-off observation of teaching". In summary, the literature includes POT models that range from accountability-based models to those that promote collaboration among teachers, and even models that aim to transform the institution's educational realities (Gosling, 2002). These existing POT models are highly relevant, hence the analysis elaborated by Byrne et al. (2010) is presented in Table 1.

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

Table 1. Models of peer observation of teaching (Byrne et al., 2010)

Characteristic	A Evaluation model	B Development model	C Peer review model	D Peer development model
Who applies it and to whom	Senior staff observe junior staff	Educational developers observe	Colleagues observe each other	Colleague engagement
Purpose	Identifies under-performance, to confirm probation, or for use in appraisal, promotion, quality assurance, assessment	Educational developers observe practitioners -or- Expert teachers observe others within a department	Colleagues observe each other teaching	Colleagues engage in exploratory dialogue about agreed sharing
Outcome	Report/judgement	Report/action plan; pass/fail	Analysis, discussion, wider experience of teaching methods	Analysis, discussion, wider experience of practice, changes
Status of evidence	Authority	Expert diagnosis	Peer shared perception	Peer shared perception
Relationship of observer to observed	Power	Expertise	Equality/mutuality	Equality/mutuality in relation to peer development
Confidentiality	Between managers, observers, and those observed	Between observer and observed, examiner	Between observer and observed – shared within learning set	

Strathern (2000) questioned whether "a university is first and foremost an organization whose performance as an organization can be observed". Nevertheless, as Biesta (2019) suggested, peer observation of teaching (POT) can be directed exclusively towards the measurement of learning, or conversely, putting these practices at the service of dialogue, mutual construction, and ultimately, education.

Although these models are not exclusive, adherence to one trend or another could condition the type of observation instrument to be designed. Similarly, models that solely address measurement may fail to consider the observing teacher's role. We must therefore remember that POT research has generally focused on the practices of observed teachers. Yet, a growing number of authors have pointed out the importance of what the observers may also be learning (Rosselló & De la Iglesia, 2021; Torres et al., 2017). From this latter perspective, that is, the dual learning of both the observers and those being observed, or mentors and mentees (Kohut et al., 2007), the initial choice of observation tool plays a key role. While interesting systematic reviews on POT in general can be found in the literature (Carragher & McGaughey, 2016; Ridge & Lavigne, 2020), none have specifically examined the tools or instruments that were employed. Our initial research revealed a number of relevant tools that have not been used in the context of university education

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

(Burgess et al., 2021; Muijs et al., 2018), whilst other tools are exhaustive (Torres et al., 2017), and some are considerably less comprehensive (Barnard et al., 2011; Bolt, 2013).

On the other hand, as Gosling (2014) pointed out, POT has been criticized for a lack of reliability when it comes to observer judgments in two aspects: the observers themselves, and the observation tool employed. In the case of observers, planning that includes various phases must be designed prior to applying POT (Cannarozzo et al., 2019). The first phase should train observers in the use of observation tools, where they are systematically trained in order to guarantee that they understand not only the theoretical basis of the instrument, but also how it should be applied in real-world situations. In the case of the tool itself, we refer to Tenbrink's (2000) definition of the evaluation concept and summarize three steps in this conceptualization: obtaining information, formulating judgments, and making decisions. Based on this definition, the following research questions are formed:

- What type of information allows judging a teacher and should be obtained?
- What tools can be used to obtain information to judge a teacher?
- What are the characteristics of tools used to judge a teacher?

Based on these research questions, we hypothesize that multiple peer observation instruments exist that are applicable within the higher education context. However, there has been no prior systematic and categorized review of such instruments to draw an accurate picture of their current status. In view of this and given the importance of the choice of observation tool in POT research, the aim of the current study was to conduct a systematic review of tools employed for peer observation of teaching based on PRISMA standards (Moher et al., 2009).

2. Methodology

2.1. Search Criteria

The searches performed centered on two criteria: a) peer observation of teaching b) higher education. Search terms were applied to the "Title/Keyword" fields. The AND command was used to join the words included in the concept: peer observation of teaching, whilst the OR command was used to separate search terms within each criterion. The complete list of search instructions was as follows:

> (TITLE (peer AND review AND of AND teaching) OR TITLE (peer AND evaluation AND of AND teaching) OR TITLE (peer AND feedback AND on AND teaching) OR TITLE (peer AND partner*) OR TITLE (formative AND observation AND of AND teaching) OR TITLE (peer AND observation) OR TITLE (mentoring) AND KEY (peer AND review AND of AND teaching) OR KEY (peer AND evaluation AND of AND teaching) OR KEY (peer AND feedback AND on AND teaching) OR KEY (peer AND partner*) OR KEY (formative AND observation AND of AND teaching) OR KEY (peer AND observation) OR KEY (higher AND education OR university* OR college* OR tertiary AND education) AND NOT TITLE-ABS-KEY (school OR elementary OR secondary OR middle)

Only articles written in the English language were included in the search. The exhaustive search equation is presented in Table 2.

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

Table 2. Search equation

a) Search criteria	1. POT		Educational Stage "a"		Educational Stage "b"
b) Search fields	Title AND Keywords		Keywords		Keywords
c) Search keywords	Peer AND review of teaching	OR	Higher AND education	AND NOT	School
	Peer AND evaluation of teaching		University*		Elementary
	Peer AND feedback AND on AND teaching		College		Secondary
	Peer AND partner		Tertiary AND education		Middle
	Formative AND observation AND of AND teaching				
	Formative AND observation AND of AND teaching				
	Peer AND observation				
	Mentoring				

The timeframe for the study was not limited. In order to be included, articles had to appear in either the SCOPUS or Web of Science databases. The search was conducted in duplicate so as to ensure accuracy and coverage, and followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The PRISMA statement is comprised of a 27-item checklist, covering the title, abstract, introduction, method, results, and discussion sections of studies in the literature. For the current study, we reviewed each article against each of the 27 items. Items 4 and 6, which refer to the PICO format, were of particular interest since they "Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design." For the current study, the "participants" were identified as the "educational stage," whereas the "interventions" referred to the "method," and "comparisons" were established using "self-assessment" as search criteria.

2.2. Exclusion criteria

Five exclusion criteria were defined and applied in the study:

1. Lack of a process of observation of teaching (e.g., mentoring without observation, observation of a written product, or solely collecting a teacher's perception or opinion of mentoring).
2. Absence of teacher-to-teacher observation.
3. Observation not having taken place in the higher education context (i.e., another educational stage).
4. Observation process conducted entirely online.
5. Insufficient detail provided regarding the instrument used.

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

2.3. Procedure

The searches were conducted between March and April of 2023. Three researchers performed searches in two databases, namely SCOPUS and Web of Science, using the search terms listed in Section 2.1 (see also Figure 1). An initial sample of 662 articles resulted from the first search (367 from SCOPUS and 295 from Web of Science). Duplicate articles (n = 18) were then eliminated, and four rounds of review were conducted based on the inclusion and exclusion criteria. During the first round, experts were informed about the context in which the observation tool was applied. In the second round, experts wrote comments on each tool after reading the abstract. A total of 502 papers were then excluded on the basis of exclusion criteria 1 (lack of process followed) and exclusion criteria 5 (insufficient detail on instrument). During the third round, tools deemed inappropriate for application in the higher education context were eliminated based on the comments received from peers, and a further 127 papers were excluded (76 from SCOPUS and 51 from Web of Science [WoS]). Finally, in the fourth round of analysis, the resulting 13 observation tools were scored, and the experts justified their scores in qualitative terms.

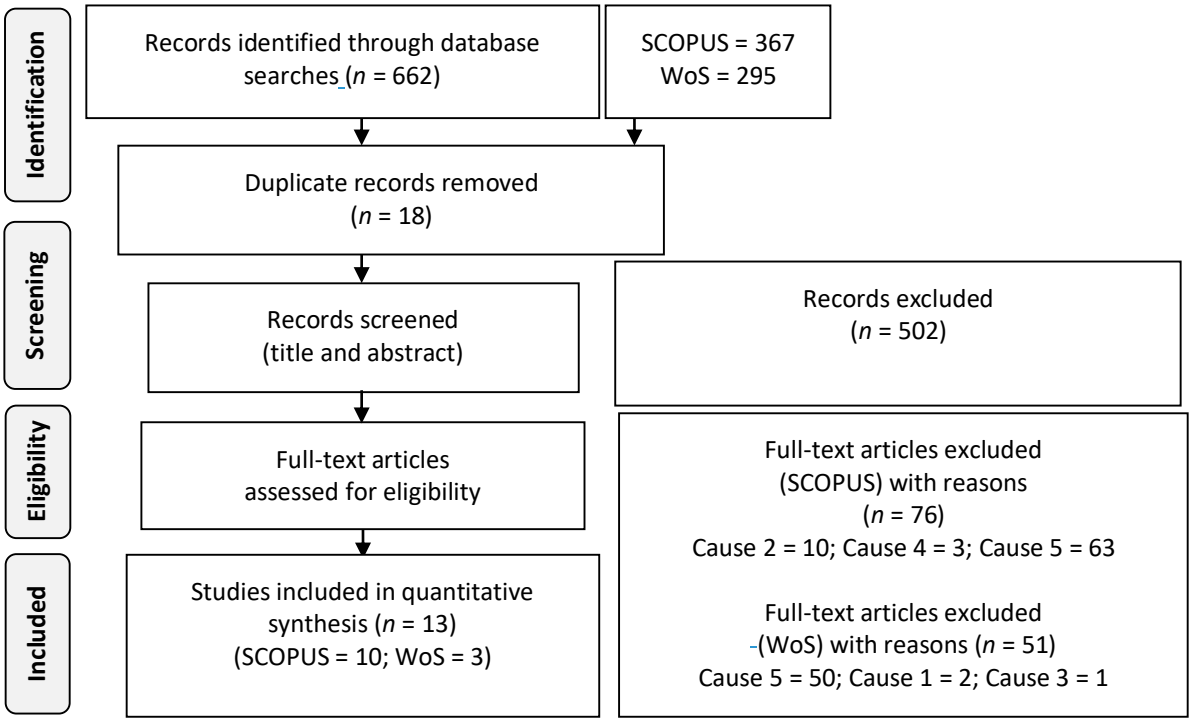

Figure. 1. Flowchart detailing the review process

2.4. Quality

The quality of research using observation tools was evaluated based on a validated instrument by López-López et al. (2019) designed for social sciences research. This tool includes 21 items arranged within eight dimensions: Cover and Summary (abstract), Introduction, Methodology, Results, Discussion, Bibliography, Appendices, and Style/Format. Each of the 21 items are evaluated according to a 5-point scale, anchored as 1 = *very low level*, 2 = *low level*, 3 = *medium level*, 4 = *medium-high level*, and 5 = *very high level*. The 21 evaluation criteria were then applied in the evaluation of the 13 studies. Three researchers independently applied the López-López et al. (2019) tool (see Table 3).

The maximum score obtainable from the tool is 105 points. The quality indicators for the studied articles were as follows: (1) mean methodological quality score for the 13 selected articles was 87.76%; (2) 11 articles scored between 80 and 105 points (excellent methodological quality);

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

(3) two articles scored between 60 and 79 points; and (4) zero articles scored below 60 points (see Table 3).

Table 3. Quality score of studies

	Items																					
	Abstract				Introduction			Methodology				Results			Discussion			B	A	Style / Format		
Authors	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Total points
Amrein-Beardsley & Popp (2012)	3	5	4	4	5	5	4	1	5	5	5	4	4	4	3	2	1	5	5	5	5	84
Bell & Mladenovic (2008)	4	5	4	5	2	5	5	5	3	5	5	5	4	5	4	5	2	5	5	5	3	92
Bolt (2013)	5	4	4	5	3	5	4	3	5	5	5	5	5	4	4	5	3	5		3	5	89
Cannarozzo et al. (2019)	5	5	4	2	5	5	5	2	4	3	5	5	5	5	4	2	5	5	5	2	4	80
Carbone et al. (2015)	5	0	5	5	4	4	4	4	5	4	0	5	5	0	5	5	4	4	4	4	5	81
Cosh (1998)	5	5	0	5	5	1	1	4	4	2	5	5	5	5	0	5	5	1	1	4	4	71
García et al. (2017)	5	4	5	5	2	5	5	5	4	4	5	5	5	4	5	5	2	5	5	5	4	97
Georgiou et al. (2018)	5	5	4	3	2	5	5	5	5	5	1	5	5	5	4	3	2	5	5	5	5	93
Hassel et al. (2020)	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	104
Rabada-Rice & Scott (1986)	5	1	3	4	2	3	2	3	1	2	1	5	5	1	3	4	2	3	2	3	1	60
Servilio et al. (2017)	5	5	5	5	5	5	5	5	5	5	1	5	5	5	5	5	5	5	5	5	5	100
Sullivan et al. (2012)	4	5	5	5	4	4	5	5	5	4	5	5	4	5	5	5	4	4	5	5	5	98
Torres et al. (2017)	4	5	5	5	3	5	5	5	4	4	1	5	4	5	5	5	3	5	5	5	4	92

B: Bibliography; A: Appendices

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

2.5. Analysis

Based on the postulates of López-Noguero (2002), a "content analysis" exercise was conducted in order to make an adequate approach to the 13 observation tools results. The content of each tool was analyzed on the basis of four variables: country, validation, observation, and feedback.

3. Results and Discussion

Based on the exclusion criteria detailed in the Methodology section, the aim of the current study was to systematically review peer observation instruments employed in university teaching. The findings revealed a total of 13 instruments which were then analyzed according to four variables: country, validation, observation, and feedback.

3.1. Country

Over half the instruments were designed by researchers from universities in two countries (see Table 4); the United States (n = 4) and Australia (n = 4). The other five instruments were originated in four European countries: The United Kingdom (n = 2), Italy (n = 1), Switzerland (n = 1), and Portugal (n = 1). Regarding the tradition of peer observation of teaching in Australia, it is worth taking note of the systematic review conducted by Johnston et al. (2022). Teaching practices are so deeply entrenched in Australia that their review found a sufficient critical mass to categories 19 studies into: a) organizational factors (disciplinary context, program sustainability, collegiality, and leadership); b) program factors (program design, basis of participation, observation, feedback, and reflective practice); and, c) individual factors (experience and participants' perceived development requirements). The extent of POT in Australia is manifest in works such as that of Bell and Cooper (2013), which explored the experiences of four Associate Deans of Learning and Teaching at a research-intensive Australian university. The Anglo-Saxon tradition of this teaching practice is clear since 10 of the 13 works found were from English-speaking countries (United States, United Kingdom, and Australia). The work of Wingrove et al. (2018) established differences between the models employed in the United Kingdom and in Australia. For Australia, "quality assurance and measurement imperatives occupy a prominent place in higher education discourses, with performativity through continuous improvement in learning and teaching now central to the very practice of learning and teaching itself" (Wingrove et al., 2018. Sachs and Parsell (2013) earlier claimed that in Australia, "peer review is neither systematically supported nor generally perceived to be a high-quality developmental activity". On the other hand, a consolidated model of peer observation of teaching is known to exist in both England and North America.

3.2. Validation

The existence of validation processes in the development of instruments advocated the need to analyze this as a variable (see Table 4). The results of the review showed that three of the 13 instruments had undergone a process of validation: the Reformed Teaching Observation Protocol (RTOP= by Amrein-Beardsley and Popp (2012); the instrument for the observation of "Project Based Learning" by García et al. (2017); and the peer observation tool introduced by Rabada and Scott (1986). The reliability and validity of RTOP were investigated within the higher education context and through examination of what participants' perceived having learned during their faculty/peer evaluation process. In the case of García et al.'s (2017) instrument designed for the observation of "Project-Based Learning," the instrument's first version was pretested and completed by four peer observers as they watched two videorecorded PBL sessions. Two videotaped tutors also completed the instrument and were then interviewed to gather their suggestions and comments; after which, certain adjustments in the formulations were applied as considered necessary. Regarding the tool designed by Rabada-Rice and Scott (1986), a group of experienced teachers deliberated on the suitability of 25 initial items, which were later reduced to 10 following a period of discussion.

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

In this regard, content validity was defined as the quality that ensures that a set of items are representative of the behavioral domain to be measured (Moscoso et al., 2003). This representativeness is usually evaluated via subjective expert opinion, which may sometimes be quantified through the use of algorithms such as the Aiken coefficient (Aiken, 1980). Therefore, the use of content validity or any other procedure that systematizes instrument design offers a level of methodological assurance that the instruments developed are both reliable and valid tools.

Table 4. Characteristics of the "observation" dimension of ETP instruments

Instrument				Observation		
Author/s	Name	Country	Val.	Dimensions	Items	Scale
Amrein-Beardsley & Popp (2012)	Reformed Teaching Observation Protocol (RTOP)	USA	Yes	1. Lesson design 2. Propositional knowledge 3. Procedural knowledge 4. Classroom culture: Interactions 5. Classroom culture: Student-teacher relationships	25	From 0 = *Not observed* To 4 = *Very descriptive of observed behavior*
Bell & Mladenovic (2008)	Peer observation proforma	Australia	No	No dimensions	16	SD = *Strongly disagree* D = *Disagree* N = *Neutral* A = *Agree* SA = *Strongly agree*
Bolt (2013)	Reflection teaching	Australia	No	1. Engagement 2. Knowledge 3. Delivery 4. Respect 5. Support 6. Connection	6	None
Cannarozzo et al. (2019)	Project mentor	Italy	No	1. Lesson management and organization 2. Teaching abilities and competencies 3. Interaction with students 4. Learning environment	24	From 1 To 4

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

Instrument				Observation		
Author/s	Name	Country	Val.	Dimensions	Items	Scale
Cosh (1998)	Observation feedback form	UK	No	Part A: What was learned Part B: Further action intended Part C: Suggested topics of interest	3	None
Drew et al. (2015)	PRO-Teaching	Australia	No	No dimensions	10	None
García et al. (2017)	Instrument for observation of "Project-Based Learning" facilitation	Switzerland	Yes	1) Tutorial: a) Problem analysis b) Self-directed learning c) Group dynamics 2) Report: a) Problem synthesis b) Discussion of group process c) Group dynamics	24	From 1 = *Makes learning uncertain* To 4 = *Optimally promotes learning*
Georgiou et al. (2018)	Teaching peer review Reviewers form	Australia	No	1. Objectives (stated/achieved) 2. Notes on objectives 3. Post-lecture discussion 4. General notes	0	None
Hassel et al. (2020)	Peer observation of small group/clinical	USA	No	1. Initiating the session 2. Presence 3. Ensuring interaction and active learning 4. Other factors contributing to effective clinical teaching and learning 5. Content and clarity 6. Closing the session	27	None
Rabada-Rice & Scott (1986)	Peer evaluation tool	USA	Yes	a) Course development b) Group participation	10	Two categories: C = Continue I = Improve

Instrument				Observation		
Author/s	Name	Country	Val.	Dimensions	Items	Scale
Servilio et al. (2017)	Peer observation form	USA	No	1. Content 2. Organization 3. Interaction 4. Verbal and nonverbal communication 5. Use of media	33	From 1 = *Poor* To 5 = *Excellent*
Sullivan et al. (2012)	Feedback form	UK	No			From 1 = *Weak* To 5 = *Strong*
Torres et al. (2017)	Peer observation of teaching	Portugal	No	1. Class structure 2. Class organization 3. Class climate 4. Content 5. Teacher attitudes 6. Other considerations	35	From 1 = *Weak* To 5 = Strong

3.3. Observation

The number of items in each tool, their grouping (or lack of grouping) into dimensions, and the response format were each aspects that were analyzed in terms of the information observed and how that information was organized for each tool (see Table 3). Regarding the number of items, the tools ranged from being highly unstructured and with just a few items to more exhaustive instruments. The least structured instruments were those developed by Cosh (1998), Drew et al. (2015), and Sullivan et al. (2012), whilst Georgiou et al.'s (2018) instrument did not include any items or dimensions.

The observation process starts with a highly detailed formulation of objectives linked to teaching. Subsequently, observation sheets guide a qualitative reflection on the teaching-learning process with respect to the formulated objectives. Along the same lines of flexibility, Cosh (1998) designed a qualitative, open instrument that consisted of three sections for the observer to complete: a) What was learned from the observation; b) Action intended to be taken (e.g., reading, staff development, further observation, and experimentation with own teaching); c) Suggested topics of interest (e.g., staff seminars, staff days, or action research). For their part, Sullivan et al. (2012) included six rating items (voice, pace, non-verbal communication, organization and preparation, use of overhead projectors, audiovisual aids, etc., and attitude). Of the four least exhaustive instruments mentioned, Drew's (2015) PRO-Teaching presented a revised version of an earlier tool. The resulting definitive version consisted of 10 largely open questions such as "Does the teacher clearly define explicit, realistic, and challenging yet achievable aims and learning objectives?" or "Does the teacher reveal a scholarly approach to teaching and seek to improve teaching performance?"

Moreover, a number of tools included multiple observation items such as those by Torres et al. (2017) (35 items), Servilio et al. (2017) (33 items), Hassel et al. (2020) (27 items), and Amrein-Beardsley and Popp's (2012) RTOP (25 items). In the case of the tool developed at the University of

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

Porto (Portugal) by Torres et al. (2017), the instrument's 35 items were structured into six dimensions (Class structure, Class organization, Class climate, Content, Teachers' attitude, and Other considerations). For its part, the instrument created by Servivlio et al. (2017) from Monmouth University (United States) groups its 33 items into five dimensions (Content, Organization, Interaction, Verbal and nonverbal communication, and Use of media). "Peer observation-clinical/small group teaching" was an observation grid designed by Hassel et al. (2020) at Colorado State University (United States) and used five dimensions to group its 27 items.

Finally, among the tools with a higher number of items, Amrein-Beardsley and Popp (2012) from Arizona State University (United States) organized the 25 questions of the RTOP instrument into four dimensions (Lesson design, Propositional knowledge, Procedural knowledge, and Classroom Culture: Interactions).

Inclusion of an item that addressed the objectives of the session under observation was found to be the most repeated item (eight of the 13 instruments) within the observation section (Bell & Mladenovic, 2008; Bolt, 2013; Cannarozzo et al., 2019; Drew et al., 2015; Georgiou et al., 2018; Hassel et al., 2020; Rabada-Rice & Scott, 1986; Torres et al., 2017).

In terms of the ideal number of items for a POT tool to consist of, we do not believe that such a number exists. However, it may be said that instruments that contain many items may risk complicating the observer's task, since they will require observers to assess many different aspects within a short period of time. Furthermore, when these tools are designed, it is essential to identify who the observers are likely to be in terms of their professional experience and job role. In the case of teacher observers, not all are likely to share the same pedagogical knowledge, and POT experiences may also be multidisciplinary (Torres et al., 2017). Thus, regardless of the number of items that an observation tool consists of, it is crucial that appropriate training programs are designed and implemented for current and future observers (Cannarozzo et al., 2019). This type of training should involve understanding the purpose of the tool in question (e.g., educational, accountability), having sufficient knowledge of the tool's items, and ensuring reliability when coding the observed behavior by referring back to the questions/items of the instrument itself.

Finally, the response format was another variable analyzed across the 13 tools reviewed. Response formats are conditioned by the number of instrument items. Tools with few questions commonly have an open response format (Bolt, 2013; Cosh, 1998; Drew et al., 2015; Georgiou et al., 2018; Rabada-Rice & Scott, 1986). On the other hand, tools with numerous items tend to have response formats with four or five options (Amrein-Beardsley & Popp, 2012; Cannarozzo et al., 2019; García et al., 2017; Servilio et al., 2017; Torres et al., 2017). Sensitivity is understood as the quality that allows for differentiation in the range of possible answers. While response formats with few options (e.g., two options) do not allow for much differentiation, it should also be noted (for tools with closed response formats) that human discrimination capacity is limited (Goñi et al., 2003). On occasion, an odd number of responses could induce the so-called central tendency error. Additionally, when POT procedures are formative in nature (beyond being designed solely for accountability), it seems appropriate to include sections where qualitative comments and clarifications can be made as closed coding response options.

3.4. Feedback

Another important aspect analyzed was whether instruments provided a section to analyze and reflect upon the observed behaviors (see Table 5). Examples of such are the inclusion of sections such as "strengths," "weaknesses," or "comments" in a developed tool. The goal being to provide feedback to the observed teacher and, ultimately, to improve the teaching and learning process. It is worthy of note that four of the 13 instruments evaluated only included an observation section, without any specific section attributed to feedback. Of the remaining instruments, some included all three aspects of "strengths," "weaknesses," and "comments" in the feedback section (Bolt,

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

2013; Cannarozzo et al., 2019; Georgiou et al., 2018), while others only included a section for "comments" or "suggestions for improvements" (Cosh, 1998; Hassel et al., 2020; Rabada-Rice & Scott, 1986; Torres et al., 2017).

If we understand POT as a phenomenon based on a critical (non-technical) paradigm and with a transformative objective, observation tools should include a specific section on feedback. A learning-oriented evaluation closes its cycle if the observed person receives adequate and valuable feedback, and if that analysis is executed horizontally (not vertically) between the observer and the observed. This type of POT model is classified under the theoretical categorization that Peel (2005) defined as "D3," that Gosling (2002) defined as the "collaborative model," and that Byrne et al. (2010) classified as the "peer development model." This model can also be said to be based upon constructivist critical teacher training (Mutlu-Gülbak, 2023).

One limitation of the current study is that only tools designed for the higher education context were reviewed. Other interesting POT tools exist in the literature that were designed for other educational stages. Also, since the COVID-19 pandemic, online learning has intensified as a more commonplace educational practice (Noor & Md Isa, 2023; Strelchuk et al., 2023; Sultoni & Gunawan, 2023); hence, another limitation of the current study is that the review excludes the influences of the pandemic on newer forms of learning.

Table 5. Characteristics of the "feedback" dimension of the POT instruments

Instrument				Feedback		
Author/s	Name	Country	Val.	Strong	Weak	Opened Notes
Amrein-Beardsley & Popp (2012)	Reformed Observation Teacher Protocol (ROTP)	USA	Yes	No	No	Yes
Bell & Mladenovic (2008)	Peer observation proforma	Australia	No	No	No	No
Bolt (2013)	Reflection teaching	Australia	No	Yes	Yes	Yes
Cannarozzo et al. (2019)	Project Mentor	Italy	No	Yes	Yes	Yes
Carbone et al. (2015)	PRO-Teaching	Australia	No			
Cosh (1998)	Observation feedback form	UK	No	No	No	Qualitative
García et al. (2017)	Instrument for observation of "Project-Based Learning" facilitation	Italy	Yes	No	No	No
Georgiou et al. (2018)	Peer Review of Teaching: Reviewers Form	Australia	No	Yes (Post-lecture discussion)	Yes (Post-lecture discussion)	Yes

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

Instrument				Feedback		
Author/s	Name	Country	Val.	Strong	Weak	Opened Notes
Hassel et al. (2020)	Peer Observation of Small Group/Clinical	USA	No	No	No	Yes "Additional comments"
Rabada-Rice & Scott (1986)	Peer Evaluation Tool	USA	Yes	No	Yes "Suggestions for improvement"	No
Servilio et al. (2017)	Peer observation forms	USA	No	Yes	Yes	No
Sullivan et al. (2012)	Feedback form	UK	No	No	No	No
Torres et al. (2017)	Peer observation of teaching	Portugal	No	No	No	Yes Appreciation of joint reflection

4. Conclusion

In this work, we conducted a systematic review of tools used for the peer observation of teaching (POT) in the higher education context. A total of 13 POT instruments were identified based on inclusion and exclusion criteria. A majority of these tools (n = 8) were developed by universities in two countries, namely the United States and Australia. Only three of the resulting instruments in the review included some form of validation process in their design. Thus, we believe that any instrument should undergo systematic processes as part of its design phase, even if the tool is to be directed towards learning-oriented evaluation rather than for the purpose of bureaucratic quality assurance. A recent systematic review published by Nuis et al. (2023) also agreed with this conclusion.

Moreover, we deem it necessary that teachers be involved in the design of instruments that are aimed towards observing the behaviors of teachers in the classroom. In can be said that teachers have long been evaluated without any involvement in the process, and the benefits of such a limited process are therefore doubtful. Furthermore, the number of items in the instruments reviewed ranged from three up to 35 questions. From our perspective, excessive numbers of questions could make the observation exercise overly complex for the observer, unless the items are purposefully distributed across several observed sessions. In any case, the actual appropriate number of questions should correspond to that deemed by teachers themselves as being essential to observing the teaching process. Hence, it is crucial that teachers help to design observation tools that may then be implemented to observe their performance behaviors.

Understanding observation from a critical viewpoint, we believe that in the case of closed-response formats, tools should include fields in which observers may add qualitative comments that can later be used to deepen the understanding of the record and to improve the feedback quality of the observer-observed encounter.

Finally, if POT is directed towards the transformation of not only technical but also human and educational realities in institutions and people, we believe that POT instruments

https://doi.org/10.22521/edupij.2024.131.6 Published online by Universitepark Press

Declarations

Author Contributions. G.G.C., F.M.O.S.: Literature review and conceptualization; F.M.O.S., J.A.D.M., G.G.C.: Methodology, data analysis; F.M.O.S., J.A.D.M.: Writing manuscript; F.M.O.S., G.G.C., J.M.C.B.: Review-editing; F.M.O.S., J.M.C.B.: Original manuscript preparation. All authors have read and approved the published final version of the article.

Conflicts of Interest. The authors declare no conflict of interest.

Funding. This research was funded by the State Research Agency under the V Own Research and Transfer Plan 2018-2020 of the Universidad Pablo de Olavide (Reference: PPI2201).

Ethical Approval. Ethical approval and participation consent were not required for this study.

Data Availability Statement. Data supporting the results presented in this study were sourced from the SCOPUS and Web of Science databases.

Acknowledgments. The authors would like to thank the Universidad Pablo de Olavide for supporting and funding this research.

References

Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, *40*(4), 955-959. https://doi.org/10.1177/001316448004000419

Amrein-Beardsley, A., & Popp, S. E. O. (2012). Peer observations among faculty in a college of education: Investigating the summative and formative uses of the Reformed Teaching Observation Protocol (RTOP). *Educational Assessment, Evaluation and Accountability*, *24*(1), 5-24. https://doi.org/10.1007/s11092-011-9135-1

Barnard, A., Croft, W., Irons, R., Cuffe, N., Bandara, W., & Rowntree, P. (2011). Peer partnership to enhance scholarship of teaching: A case study. *Higher Education Research and Development*, *30*(4), 435-448. https://doi.org/10.1080/07294360.2010.518953

Bell, A., & Mladenovic, R. (2008). The benefits of peer observation of teaching for tutor development. *Higher Education*, *55*(6), 735-752. https://doi.org/10.1007/s10734-007-9093-1

Bell, M., & Cooper, P. (2013). Peer observation of teaching in university departments: A framework for implementation. *International Journal for Academic Development*, *18*(1), 60-73. https://doi.org/10.1080/1360144X.2011.633753

Biesta, G. (2019). What is the Educational Task? Arousing the Desire for Wanting to Exist in the World in a Grown-up Way. *Pedagogía y Saberes*, *50*, 51-61. https://doi.org/10.17227/pys.num50-9498

Bolt, S. (2013). Closing the Loop with Collegiate Conversations in an Australian Voluntary Peer Review of Teaching Program. *International Journal on Teaching and Learning in Higher Education*, *19*(3), 1-15. https://doi.org/10.18848/1447-9494/CGP/v19i03/48643

Brooks, C., Carroll, A., Gillies, R. M., & Hattie, J. (2019). A matrix of feedback for learning. *Australian Journal of Teacher Education*, *44*(4), 14-32. https://doi.org/10.14221/ajte.2018v44n4.2

Brooks, C., Huang, Y., Hattie, J., Carroll, A., & Burton, R. (2019). What Is My Next Step? School Students' Perceptions of Feedback. *Frontiers in Education*, *4*, Article 96. https://doi.org/10.3389/feduc.2019.00096

Burgess, S., Rawal, S., & Taylor, E. S. (2021). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. *Journal of Labor Economics*, *39*(4), 1155-1186. https://doi.org/10.1086/712997

Byrne, J., Brown, H., & Challen, D. (2010). Peer development as an alternative to peer observation: A tool to enhance professional development. *International Journal for Academic Development*, *15*(3), 215-228. https://doi.org/10.1080/1360144X.2010.497685

Cannarozzo, M., Gallo, P., Lo Coco, A., Megna, B., Musso, P., & Scialdone, O. (2019). The Peer Observation: "Mentore" Project at University of Palermo. *Knowledge Management and Organizational Learning*, *8*, 237-248. https://doi.org/10.1007/978-3-030-29872-2_14

Carbone, A., Ross, B., Phelan, L., Lindsay, K., Drew, S., Stoney, S., & Cottman, C. (2015). Course evaluation matters: improving students' learning experiences with a peer-assisted teaching programme. *Assessment and Evaluation in Higher Education*, *40*(2), 165-180. https://doi.org/10.1080/02602938.2014.895894

Carragher, J., & McGaughey, J. (2016). The effectiveness of peer mentoring in promoting a positive transition to higher education for first-year undergraduate students: A mixed methods systematic review protocol. *Systematic Reviews*, *5*(1), Article 68. https://doi.org/10.1186/s13643-016-0245-1

Cosh, J. (1998). Peer observation in higher education – A reflective approach. *Innovations in Education and Teaching International*, *35*(2), 171-176. https://doi.org/10.1080/1355800980350211

Drew, S., Klopper, C., & Nulty, D. (2015). Defining and developing a framework for the peer observation of teaching. In C. Klopper & S. Drew (Eds.), *Teaching for Learning and Learning for Teaching* (pp. 13-34). Sense Publishers. https://doi.org/10.1007/978-94-6300-289-9_2

García, I., James, R. W., Bischof, P., & Baroffio, A. (2017). Self-Observation and Peer Feedback as a Faculty Development Approach for Problem-Based Learning Tutors: A Program Evaluation. *Teaching and Learning in Medicine*, *29*(3), 313-325. https://doi.org/10.1080/10401334.2017.1279056

Georgiou, H., Sharma, M., & Ling, A. (2018). Peer review of teaching: What features matter? A case study within STEM faculties. *Innovations in Education and Teaching International*, *55*(2), 190-200. https://doi.org/10.1080/14703297.2017.1342557

Goñi, A., Esnaola, I., Ruiz De Azua, S., Rodriguez, A., & Zulaika, L. M. (2003). Autoconcepto físico y desarrollo personal: perspectivas de investigación [Physical self-concept and personal development: research perspectives]. *Revista de Psicodidáctica*, *15-16*, 7-62. https://ojs.ehu.eus/index.php/psicodidactica/article/view/156

Gosling, D. (2002, August). *Models of peer observation of teaching* [Keynote address at LTSNGC Peer Observation of Teaching Conference.

Gosling, D. (2014). Collaborative Peer-Supported Review of Teaching. In J. Sach and M. Parsell (Eds.), *Peer Review of Learning and Teaching in Higher* (pp. 13-32). Springer.

Hassel, D. M., Fahie, M., Loehr, C. V., Halsey, R. L., Vernau, W., & Gorman, E. (2020). Inter-institutional collaboration for the development of a local peer observation process to enhance teaching. *Journal of Veterinary Medical Education*, *47*(5), 555-569. https://doi.org/10.3138/JVME-2019-0093

Hattie, J., & Clarke, S. (2018). *Visible learning: Feedback.* Routledge. https://doi.org/10.4324/9780429485480

Johnston, A. L., Baik, C., & Chester, A. (2022). Peer review of teaching in Australian higher education: a systematic review. *Higher Education Research and Development*, *41*(2), 390-404. https://doi.org/10.1080/07294360.2020.1845124

Kohut, G. F., Burnap, C., & Yon, M. G. (2007). Peer Observation of Teaching: Perceptions of the Observer and the Observed. *College Teaching*, *55*(1), 19-25. https://doi.org/10.3200/CTCH.55.1.19-25

López-López, E., Tobón, S., & Juárez-Hernández, L. G. (2019). Scale to evaluate scientific articles in social and human sciences – SSAHS. *Revista Iberoamericana Sobre Calidad, Eficacia y Cambio En Educacion*, *17*(4), 111-125. https://doi.org/10.15366/REICE2019.17.4.006

López-Noguero, F. (2002). El Análisis de contenido como método de investigación [Content analysis as a research method]. *Revista de Educación*, *4*, 167-180. https://www.uhu.es/publicaciones/ojs/index.php/xxi/article/view/610

López Pastor, V. M. (1999). *Prácticas de evaluación en educación física: estudio de casos en primaria, secundaria y formación de profesorado* [Assessment practices in physical education case study in primary, secondary and teacher training] [Doctoral dissertation, University of Valladolid, Spain]. https://produccioncientifica.ucm.es/documentos/5d1df62729995204f7662e4c?lang=gl

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moscoso, S. C., Antonio Pérez-Gil, J., Pablo, F., Tello, H., & Ruiz, Á. L. (2003). Evaluación de la calidad universitaria: validez de contenido [Evaluation of university quality: validity of content. Psychotherna]. *Psicotherna*, *13*(2), 294-301. https://reunido.uniovi.es/index.php/PST/article/view/7833

Muijs, D., Reynolds, D., Sammons, P., Kyriakides, L., Creemers, B. P. M., & Teddlie, C. (2018). Assessing individual lessons using a generic teacher observation instrument: how useful is the International System for Teacher Observation and Feedback (ISTOF)? *ZDM - Mathematics Education*, *50*(3), 395-406. https://doi.org/10.1007/s11858-018-0921-9

Mutlu-Gülbak, G. (2023). Expectations for Training Mentors: Insights from a Preservice Language Teacher Education Program. *Educational Process: International Journal, 12*(2), 76-92. https://dx.doi.org/10.22521/edupij.2023.122.5

Noor, S., & Md Isa, F. (2023). Online Learning Challenges Faced by SSC-Level Learners During Pandemic: A Case of Pakistan. *Educational Process: International Journal, 12*(4), 65-77. https://dx.doi.org/10.22521/edupij.2023.124.4

Nuis, W., Segers, M., & Beausaert, S. (2023). Conceptualizing mentoring in higher education: A systematic literature review. *Educational Research Review*, *41*, Article 100565. https://doi.org/10.1016/j.edurev.2023.100565

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, *35*(1), Article 100416. https://doi.org/10.1016/j.edurev.2021.100416

Peel, D. (2005). Peer observation as a transformatory tool? *Teaching in Higher Education*, *10*(4), 489-504. https://doi.org/10.1080/13562510500239125

Rabada-Rice, F., & Scott, R. S. (1986). A peer evaluation for measuring team teaching effectiveness. *The Journal of Nursing Education*, *25*(6), 255-258. https://doi.org/10.3928/0148-4834-19860601-10

Ridge, B. L., & Lavigne, A. L. (2020). Improving instructional practice through peer observation and feedback. *Education Policy Analysis Archives*, *28*(61). https://doi.org/10.14507/EPAA.28.5023

Rosselló, M. R., & De la Iglesia, B. (2021). El feedback entre iguales y su incidencia en el desarrollo profesional docente [Feedback between equals and its impact on professional teaching development]. *Revista Complutense de Educación*, *32*(3), 371-382. https://doi.org/10.5209/rced.70173

Sachs, J., & Parsell, M. (2013). The place of peer review in learning and teaching. In J. Sachs and M. Parsell (Eds.), *Peer Review of Learning and Teaching in Higher* (pp. 1-10). Springer.

Servilio, K. L., Hollingshead, A., & Hott, B. L. (2017). Partnerships Enhancing Practice: A Preliminary Model of Technology-Based Peer-to-Peer Evaluations of Teaching in Higher Education. *Journal of Special Education Technology*, *32*(1), 23-35. https://doi.org/10.1177/0162643416681161

Strathern, M. (2000). The Tyranny of Transparency. *British Educational Research Journal*, *26*(3), 309-321. https://doi.org/https://doi.org/10.1080/713651562

Strelchuk, E. N., Kozhevnikova, M. N., & Borchenko, V. S. (2023). Blended Learning in Russian Higher Education: The Evolution of the Term in Science and Practice. *Educational Process: International Journal, 12*(1), 97-116. https://dx.doi.org/10.22521/edupij.2023.121.6

Sullivan, P. B., Buckle, A., Nicky, G., & Atkinson, S. H. (2012). Peer observation of teaching as a faculty development tool. *BMC Medical Education*, *12*, Article 26 . https://doi.org/10.1186/1472-6920-12-26

Sultoni, & Gunawan, I. (2023). Relationship between Perceived Transformational Leadership and Organizational Citizenship Behavior of Virtual Teaching During the COVID-19 Pandemic in Indonesia: The Mediating Role of Job Satisfaction. *Educational Process: International Journal, 12*(3), 56-78. https://dx.doi.org/10.22521/edupij.2023.123.3

Tenbrink, T. D. (2000). Evaluación [Evaluation]. In J. Cooper (Ed.), *Estrategias de enseñanza: (guía para una mejor instrucción)* [Teaching strategies: (guide to better instruction)] (pp. 499-558). Limusa.

Torres, A. C., Lopes, A., Valente, J. M. S., & Mouraz, A. (2017). What catches the eye in class observation? Observers' perspectives in a multidisciplinary peer observation of teaching program. *Teaching in Higher Education*, *22*(7), 822-838. https://doi.org/10.1080/13562517.2017.1301907

Wingrove, D., Hammersley-Fletcher, L., Clarke, A., & Chester, A. (2018). Leading Developmental Peer Observation of Teaching in Higher Education: Perspectives from Australia and England. *British Journal of Educational Studies*, *66*(3), 365-381. https://doi.org/10.1080/00071005.2017.1336201

About the Contributor(s)

Fernando Manuel Otero Saborido lectures in physical education at Pablo de Olavide University (Seville, Spain). His main research interests include topics related to planning, methodology, and evaluation in physical education.
Email: fmotero@upo.es
ORCID ID: https://orcid.org/0000-0002-7016-2414

José Antonio Domínguez-Montes is a teacher of physical education at the secondary school level (Seville, Spain). His main research interests include topics related to didactics in physical education.
Email: jadmef03@hotmail.com
ORCID ID: https://orcid.org/0000-0002-7207-4143

José Manuel Cenizo Benjumea lectures in physical education at Pablo de Olavide University (Seville, Spain). His main research interests include topics related to didacts in physical education and motor development.
Email: jmcenben@upo.es
ORCID ID: https://orcid.org/0000-0002-8009-4806

Gustavo González Calvo lectures in physical education at University of Valladolid (Spain). His main research interests include topics related to methodology and evaluation, and learning in physical education.
Email: gustavo.gonzalez@uva.es
ORCID ID: https://orcid.org/0000-0002-4637-0168