EARLY VIEW
doi.org/10.54337/ojs.jpblhe.v14i1.10158

Journal of
Problem Based Learning
in Higher Education

EIPA-PBL: An Embedded, Real-Time Approach to Individualised Performance Assessment in Problem-Based Learning Tutorials

A Pilot Instrument-Development Study

Colin Greengrass * | RCSI Medical University of Bahrain (RCSI-MUB), Bahrain

Abstract

Problem-based learning (PBL) tutorials are designed to make learning visible through collaborative inquiry, yet the evidence generated in these settings is often collective, distributed and difficult to attribute to individual students. Programmes must nevertheless make defensible individual judgements for feedback, remediation, progression and, in some contexts, summative assessment. This creates a persistent assessment problem: the process evidence most relevant to PBL is difficult to capture, while the scores required for institutional decision-making risk compressing collaborative performance into overly simple or performative measures. Existing approaches, including peer ratings, group products, post-session tutor judgements and lengthy item-based instruments, may provide limited individualisation or impose substantial rater burden. This pilot instrument-development study reports the design and initial evaluation of EIPA-PBL, an Embedded Individualised Performance Assessment approach for PBL tutorials. EIPA-PBL uses a schematic seating map and a small set of observable indicators to record individual learning-relevant behaviours during tutorials. Individualisation refers to the generation of indi-

* Corresponding author:
Colin Greengrass, Email: cgreengrass@rcsi-mub.com

Received 4 May 2025 | Accepted 22 June 2026 | Available online in Early View 3 July 2026

vidual evidence profiles within a shared indicator framework, rather than the use of different criteria for different students. Data from 37 medical students across five PBL modules yielded 43 student-module enrolment records and approximately 7,000 coded indicator events. Intraclass correlation coefficients showed indicator-specific cross-session patterns, while exploratory factor analysis suggested coherent clustering of background knowledge, ideas and minor contributions, with major contributions and questions requiring refinement. Student feedback provided preliminary acceptability evidence. The study broadly supports feasibility and preliminary construct exploration, but inter-rater reliability, response-process evidence and consequences research remain necessary. EIPA-PBL is presented as a candidate framework for formative, portfolio and bounded low-stakes summative use.

Keywords: problem-based learning; individualized assessment; group work assessment; tutorial assessment; observational assessment; collaborative learning; performance indicators

Introduction

Problem-based learning (PBL) tutorials are designed around collaborative inquiry, wherein learners engage with complex and often ill-structured problems (Savery, 2006) to identify learning needs, undertake self-directed study and return to the group to test explanations, revise understanding and develop shared accounts of the problem (Barrows & Tamblyn, 1980; Dolmans et al., 2005; Hmelo-Silver, 2004). Dillenbourg (1999) highlighted that collaborative learning allows components of cognition to become externalised through articulation at multiple levels within the process, requiring students to explain their thinking to others. In this regard, tutorial participation is not peripheral to learning in PBL – it is one of the principal means by which students make developing understanding available for negotiation and revision, through explanation, questioning and the collaborative construction of problem scaffolds (Chi & Wylie, 2014; Hmelo-Silver & Barrows, 2008; Webb, 1989).

This creates a specific assessment challenge. PBL activity is fundamentally collaborative, whereas institutional decisions about progress, feedback, and remediation are invariably made at the individual-student level. Shared outputs, such as group explanations, whiteboard summaries, learning objectives or presentations, cannot easily show how each student contributed to the collaborative task (Dijkstra et al., 2016; Strijbos, 2011). In PBL tutorials, evidence of individual contribution is often scattered across brief moments of interaction, different phases of the tutorial, and both formal and informal roles.

Assessment therefore needs to recognise the group collaborative setting while still preserving evidence of each student's contribution within it.

The need for individual-level evidence is further underscored by the documented impact of assessment on participation within PBL. Czabanowska et al. (2012) noted the presence of free riders who may feign active involvement, whilst Savin-Baden (2004) observed that without assessment, students are more prone to rote memorisation, description rather than critique, and selective attendance. Van Berkel and Schmidt (2000) further demonstrated that individual commitment is the most potent determinant of achievement in PBL sessions – supporting the view that assessment within PBL may be necessary to ensure active participation.

A further difficulty is that assessment of performance in teams may be negatively impacted by approaches requiring uniform behavioural demonstration from all group members. A specific set of behaviours outlined in a rubric may lead group members to adhere to a mental checklist they feel obligated to demonstrate, potentially hindering the natural flow of collaborative activities (Bearman & Ajjawi, 2018; Torrance, 2007). Teamwork theory describes performance as a dynamic process involving differentiated and complementary transition, action and interpersonal processes varying with task phase and group need (Kozlowski & Ilgen, 2006; Marks et al., 2001; Salas et al., 2005). Within a single session, different students may advance the same collaborative task through entirely different contributions, none of which need be displayed uniformly.

Various methods of assessment of performance in PBL sessions have been documented within review articles (MacDonald, 2005; Nendaz & Tekian, 1999), with process-oriented instruments most commonly utilising post-session tutor ratings or peer evaluation (Segers & Dochy, 2001). TUTOTEST, for example, provides a structured approach to tutorial assessment, but its 44-item format illustrates the burden generated when tutors must judge several students across multiple domains (Hébert & Bravo, 1996). More recent instruments such as the Collaborative Learning Development Exercise (CLeD-EX) provide structured tutor feedback on collaborative learning behaviours in medical students after the session (Pervaz Iqbal et al., 2020), and observational tools have been developed to code instructor and student behaviours in classroom and PBL contexts (Alimoglu et al., 2014). Nevertheless, many of these approaches remain post-session or retrospective, requiring tutors to translate complex interaction into global ratings after the event. Eva (2001) described potential psychometric weaknesses arising from tutors being expected to evaluate numerous qualities simultaneously for multiple students, with assessors prone to recall bias and anchoring to memorable events from tutorials that may have taken place

several weeks earlier (Eva et al., 2007; Gingerich et al., 2014; Norman et al., 2006; Tavares & Eva, 2013).

In this study, a form of assessment was designed to address these limitations by recording performance variables for individuals in real time, without assuming that all students must demonstrate the same indicators in the same way. EIPA-PBL is an embedded, real-time observational assessment approach using a schematic representation of the tutorial seating arrangement and a limited shared repertoire of performance indicators. It enables each group member to contribute within naturally assumed roles and collaborative task demands, while preserving individual-level evidence within the collective activity. The construct of interest is observable individual performance within collaborative PBL taskwork – the behaviours through which a student contributes to the group's work of constructing, testing and revising explanations of a problem. EIPA-PBL is therefore intended to record observable performance within the tutorial process, not to provide a comprehensive measure of student learning, motivation or personal qualities as a team member.

In practical terms, individualised assessment means that the tutor records what each student actually does during the tutorial, using the same set of indicators for everyone, and then interprets each student's pattern of evidence in relation to their role, opportunities and contribution to the group task. It does not mean that different students are assessed by different standards; rather, it means that students may demonstrate valuable performance through different combinations of observable behaviours. Fairness is therefore understood as assessment of the same underlying construct through a transparent interpretive framework, rather than as a requirement that all students display the same behaviours.

This pilot study developed EIPA-PBL and explored its feasibility, cross-session indicator behaviour and preliminary internal structure during live PBL tutorials. The study addressed three development questions: (1) Could EIPA-PBL be completed during live PBL sessions while still allowing tutorial facilitation? (2) How did its indicators behave across repeated tutorial sessions, given the expected context-sensitivity of PBL performance? (3) What preliminary internal structure was suggested by exploratory factor analysis, and what did this imply for instrument refinement?

Theoretical framing

Externalised cognition and observable tutorial performance

The theoretical basis for observing performance in PBL is that tutorial discourse externalises parts of the learning process. Collaborative learning requires students to explain, justify, question, compare and revise ideas, making aspects of cognition visible to peers and tutor (Dillenbourg, 1999; Hmelo-Silver & Barrows, 2008; Mercer, 2000). These explanations support elaboration and restructuring of knowledge, whilst peer interaction prompts learners to identify gaps and refine understanding (Slavin, 1996; Webb, 1989). The ICAP framework distinguishes passive, active, constructive and interactive modes of engagement, providing theoretical support for the view that interactive and constructive behaviours offer stronger evidence of learning activity than attendance or silent exposure alone (Chi & Wylie, 2014). Crucially, ICAP's argument rests on overt, observable student behaviours as proxies for cognitive engagement modes – a position which directly supports the validity of event-level behavioural recording as meaningful evidence of learning-relevant activity.

This does not mean that all learning is visible. Silent listening, private reflection and internal cognitive reorganisation may be educationally important. The construct claim for EIPA-PBL is therefore deliberately bounded: the instrument records learning-relevant behaviours that occur within tutorial activity, not the totality of student learning.

Distributed cognition and individualised assessment within a shared framework

PBL tutorial activity can be understood as a distributed cognitive system in which understanding is constructed across participants, artefacts (whiteboard, notes, concept maps) and shared talk (Hutchins, 1995; Stahl, 2006). Within such a system, individual contributions are best interpreted relationally: their significance depends on the state of the group task, the role being occupied, and the contribution's effect on the developing shared explanation. This is consistent with team-process theory, which treats collective performance as phase-sensitive and role-differentiated rather than behaviourally uniform (Kozlowski & Ilgen, 2006; Marks et al., 2001; Salas et al., 2005). If assessment is interpreted as requiring every student to demonstrate every behaviour, it risks imposing behavioural uniformity on a process that functions through differentiation. The implication for EIPA-PBL is that the shared indicator framework provides a common language for observation, while each student's profile records the particular pathway through which they contributed to the

group task. This addresses the long-recognised "individual within the collective" problem in collaborative learning assessment (Strijbos, 2011) and aligns with collaborative-learning design frameworks that embed assessment within the collaborative task to support equitable participation and valid interpretation of individual roles (de Hei et al., 2016), as well as sustainable assessment approaches emphasising longitudinal, profile-based evidence of learner development (Boud & Soler, 2016).

Rubrics, performativity and the need for a bounded claim

A structured indicator set inevitably resembles a rubric in some respects, and it is important to acknowledge both the utility and the limitations of this parallel. Indicator sets provide a common language for judgement and can support transparency and consistency (Jonsson & Svingby, 2007; Reddy & Andrade, 2010); the claim for EIPA–PBL is not that it avoids structure, but that it uses structure differently. It is an embedded recording schema for contemporaneous evidence capture, not a post–hoc checklist requiring every student to perform every descriptor.

This distinction matters because preset criteria can have unintended educational consequences. Although criteria can support transparency and consistency, they may also narrow attention, encourage strategic compliance and create the misleading impression that complex judgement has been reduced to objective measurement (Bearman & Ajjawi, 2018; Sadler, 2009; Torrance, 2007). Thus, assessed participation can become public performance, particularly where students perceive that visible behaviour is being continuously monitored (Macfarlane, 2015). EIPA-PBL does not remove this risk. However, its indicators are intended to record contributions that materially support the tutorial task, rather than visible activity for its own sake. In this sense, the instrument may direct students towards productive forms of participation, such as questioning, explanation, synthesis, use of background knowledge and responsibility for group process, while making purely performative behaviour less useful unless it contributes to collaborative inquiry.

Most performativity research, however, examines post-hoc rubric-driven assessment; live observational assessment is a comparatively under-developed sub-literature, and the consequences of this design choice require specific examination in real-time tutorial conditions. EIPA-PBL should therefore be understood as a tool that structures tutor observation and feedback while acknowledging that observation may itself influence the behaviour being observed.

Development of EIPA–PBL

Design principles

The development of EIPA–PBL was driven by four practical requirements. The instrument needed to capture evidence during the tutorial rather than relying on memory after the event; employ a manageable number of indicators so that the tutor could continue facilitating whilst recording; document individual evidence within the spatial and social arrangement of the tutorial; and support individualised feedback without requiring uniform behavioural performance from all group members.

The cognitive demand placed on the tutor was central to the design rationale. During a PBL tutorial, tutors must facilitate inquiry, monitor group dynamics, attend to content accuracy, support psychological safety and observe individual learners (Hmelo-Silver & Barrows, 2006). An embedded assessment instrument must therefore be usable under conditions of limited working memory and ongoing interactional demand (Cowan, 2010; Tavares & Eva, 2013). Cognitive load theory further suggests that instructional or assessment designs should minimise unnecessary processing demands so that limited cognitive resources can be directed toward the task itself (van Merriënboer & Sweller, 2010). This is particularly important in rater–based assessment, where judgement is shaped by attention, memory, interpretation and workload (Gingerich et al., 2014; Tavares & Eva, 2013). Criterion efficiency is therefore essential: each additional indicator imposes a continuous cognitive cost while the tutorial unfolds. For this reason, EIPA–PBL uses a small indicator set and a schematic spatial layout. Spatial representations can support memory and interpretation by organising semantic and spatial relations in a single visual field (Kulhavy & Stock, 1996; Scott & Schwartz, 2007).The seating map reduces the need to translate student identity into rows in a table and provides a visual organiser for linking observed events to individual students.

Selection of performance indicators

Indicators were selected to represent observable manifestations of collaborative cognition and taskwork in PBL, whilst remaining feasible for real–time recording. The intention was not to measure every desirable interpersonal quality; rather, indicators were chosen to represent behaviours through which students make thinking visible, advance problem analysis, support group process and contribute to task completion. These categories should be understood as event classes rather than broad personality traits or fixed competencies, selected because each represents a plausible and observable route through which a student may contribute to collaborative inquiry. The criteria utilised are as follows:

Responsibility refers here to observable process–regulation behaviours rather than to a general personal quality. It includes actions that organise, direct or stabilise the group's work, such as guiding the group through problem analysis, prompting the next step, supporting the leader or scribe, keeping the group aligned with the learning task and helping the group transition from discussion to identification of self–directed learning needs. In this sense, responsibility is used as practical shorthand within the EIPA–PBL schematic for process leadership or task regulation. It should not be interpreted as a judgement that a student is generally "responsible", nor should isolated or opportunistic gestures be over–valued where they do not materially support the group's taskwork. This interpretation is grounded in team–process theory and self–regulated learning, in which effective collaboration requires planning, monitoring and regulation (Johnson & Johnson, 2005; Marks et al., 2001).

Background knowledge records the use of relevant factual or conceptual knowledge during the tutorial. The PBL process depends on the activation of prior knowledge and elaboration through discussion (Schmidt, 1983; Schmidt et al., 2011); where students introduce relevant facts, mechanisms, definitions or clinical relationships, they provide resources which the group can test, connect and use in constructing an explanation. This indicator therefore represents an important pathway through which individual preparation and prior knowledge become visible within the group.

Ideas and creative thinking records generative reasoning. In PBL, students do not merely retrieve facts; they propose hypotheses, construct explanations, identify alternative interpretations and build problem representations (Hmelo-Silver, 2004; Nijstad & Stroebe, 2006). This criterion records acts that redirect the group's reasoning, propose new explanatory frames, challenge an erroneous line of thought, support development of a problem scaffold or produce a useful representation on the whiteboard.

Questions are recorded because questioning is central to inquiry within PBL. A question can reveal a knowledge gap, test an assumption, request clarification, challenge a claim or direct the group toward deeper reasoning (Chin & Osborne, 2008; King, 1994). This criterion aligns with metacognitive monitoring, in which learners use questions to check and regulate their comprehension (Schraw & Dennison, 1994), and with elaborative interrogation, where generating explanatory questions promotes deeper learning.

Minor contributions and *major contributions* were included for a pragmatic reason rooted in the realities of real–time observation. As Gingerich et al. (2014, p. 1058) note, "there can be no such thing as 'objective' observation of performance"; raters must, in real time, decide how to categorise events whilst subsequent events continue to occur. *Minor contribution* (notation – open circle) therefore

serves as a low–inference record of brief or difficult–to–classify participation, preventing data loss when an event cannot confidently be coded under a more specific indicator within the available time. *Major contribution* records extended or substantial turns. These categories were not designed to replace the more specific indicators but to maintain the feasibility of embedded recording under time pressure; their later behaviour in factor analysis is therefore particularly informative for instrument refinement, as discussed below.

The schematic format and the null marker

The instrument is presented as a schematic map of the tutorial space (Figure 1). Each student is represented in a position corresponding to the seating arrangement, and the tutor marks indicators directly within the student's space on the schematic. This format affords the tutor immediate visual awareness of emerging participation patterns, including students with sparse records, dominant speakers and students not yet invited into the discussion. The schematic is therefore not merely a visual style – it is part of the instrument's cognitive and equity logic, enabling real–time awareness of participation balance across the group.

The null marker was included to distinguish between not being invited to participate and declining or not responding when directly invited. Non–participation may reflect social loafing (Karau & Williams, 1993; Latané et al., 1979), low confidence, limited preparation, language barriers, exclusion by the group (Visschers-Pleijers et al., 2006) or internal processing. The null marker is primarily diagnostic and formative: it is intended to prompt tutor attention to participation equity and to enable timely corrective intervention, rather than to function as a negative score in any summative system.

Operational definitions

In applying operational definitions, *background knowledge* is acceptable when the student introduces knowledge helping the group clarify the problem, test an explanation, identify a mechanism or decide what must be learned next; it should not reward disconnected fact–listing. *Ideas and creative thinking* is acceptable when the student produces generative movement in group understanding – proposing a hypothesis, reframing the problem, challenging a mistaken assumption, connecting previously separate points or producing a useful whiteboard representation. The terms *minor* and *major contribution* should not be understood as simple measures of value: a *minor contribution* (notation – open circle) records a brief, relevant act that supports the tutorial but cannot be sufficiently classified under a more specific indicator in real time, whilst a *major contribution* (notation – closed circle) records an extended or

substantial act which – as the factor analysis forthwith makes clear – is not always equivalent to higher–order collaborative performance.

Negative behaviour was operationalised narrowly because negative marking carries greater interpretive and ethical risk than the recording of positive task contributions. The marker was reserved for repeated or clearly disruptive observable acts that materially impeded collaborative inquiry, such as persistent interruption after redirection, dismissive responses to peer contributions, verbal dominance that prevented others from participating, or actions that repeatedly derailed the task. It was not intended to capture isolated lapses, enthusiasm-driven interruptions, anxiety-related behaviours, neurodivergent communication patterns, second-language processing or cultural differences in interaction style. The same surface behaviour may have different meanings in different contexts; negative marking should therefore be treated as a contextual judgement about behaviour within the specific tutorial task setting, not as evidence of personality, motivation or professionalism. This reflects a central principle of EIPA-PBL: indicators are operational categories for observable tutorial performance, not labels for stable personal traits.

Use of EIPA–PBL data and scoring

EIPA–PBL has formative, portfolio and bounded summative uses. Because the instrument generates indicator–specific records, it can support targeted feedback on the pattern of a student's performance; i.e. frequent use of background knowledge but limited questioning, strong responsibility for the group process but limited synthesis, or meaningful questions with few extended explanations. Such profiles are more useful for development than a single global participation score and align with sustainable, longitudinal assessment approaches (Boud & Soler, 2016).

In the implementation reported here, EIPA–PBL data contributed to a low–stakes summative score alongside other forms of assessment. Indicators were assigned weights reflecting their differing importance in the successful conclusion of activities: minor contribution (1), major contribution (2), responsibility (1), ideas and creative thinking (2), background knowledge (2), relevant question (1), negative behaviour (−1) and null response (0). The null marker did not contribute to summative totals, serving only as a diagnostic prompt. It should be noted that where a numerical summary is retained, it represents a compression of the richer evidence profile and should not be treated as the assessment in itself; the profile is the more educationally informative output. Optimal use is formative and portfolio–based, with summative use bounded and undertaken only alongside multiple other sampling points.

Methods

Context and participants

The study took place within an undergraduate medical programme in Riyadh, Saudi Arabia in which PBL formed a core curricular component. Tutorials followed a conventional PBL cycle, consistent with the Maastricht seven-jump logic (Schmidt, 1983), involving initial clarification and problem analysis, identification of learning needs, self-directed preparation between sessions, and subsequent reporting or synthesis phases. Formal roles, including group leader and scribe, were used and rotated across sessions.

Before the main study, two pilot cohorts of 10 students each were assessed to explore whether EIPA–PBL scores showed associations with external performance measures, and how these associations compared with a rubric–based assessment used locally. These pilot data were analysed correlationally and are reported as preliminary construct–relevant evidence only.

The main study included 37 medical students across five PBL modules. Six students completed two modules each, yielding 43 student-module enrolment records across six PBL sessions per module. This produced 258 student-session records in total. Groups ranged in size from 8 to 10 students. Across these records, the schematic generated approximately 7,700 coded indicator events. The dataset was nested, where multiple coded events came from the same student within the same session, and multiple sessions came from the same student and module. These coded events provided a dense observational record, but they were not treated as statistically independent observations.

Ethics, consent and feedback

The study received institutional ethics approval. EIPA–PBL was used as part of normal programme–required low–stakes assessment, regardless of whether students agreed to research participation. Consent therefore covered only the use of routinely collected EIPA–PBL data for research purposes, not participation in any additional activity. All 37 students within the author's assigned PBL groups were invited to consent to the use of their data; all provided written informed consent, and none subsequently withdrew.

Declining consent had no consequence for students' teaching, assessment or relationship with the author. Nevertheless, because the study was conducted within an assessed educational setting in which the author also held a teaching role, students' perceived freedom to decline may have been constrained. To mitigate this, students retained the right to withdraw their data after completion of the programme, when they had progressed into Year 3 Semester

2 clinical placements and were no longer being taught or assessed by the author. This right was reiterated in writing at the point of transition.

Completed EIPA-PBL schematics were used to support brief post-session feedback to students. Feedback focused on patterns of observed contribution, including use of background knowledge, questioning, ideas generation, responsibility for group process and balance of participation, rather than on isolated marks.

Student feedback on the instrument was then collected during brief post-session group discussions in the fourth week of each PBL sequence, broadly following a truncated focus-group format (Stewart et al., 2007). Discussion prompts asked whether EIPA-PBL provided reliable and timely feedback, allowed students to participate freely without consciously performing to assessment criteria, and encouraged participation. Students were also invited to identify negative aspects of the instrument.

Data analysis

Data were analysed using SPSS. Intraclass correlation coefficients (ICC, two–way mixed, average measures, absolute agreement) were calculated to examine indicator–specific patterns across the six sessions. Cross–session ICC values are not interpreted here as a simple test of behavioural stability: PBL performance is expected to vary with case content and prior knowledge (Norman et al., 2006; Schmidt et al., 2011), with formal role assignment (Marks et al., 2001), and with tutorial phase and group dynamics (Hmelo-Silver & Barrows, 2008; Visschers-Pleijers et al., 2006). ICCs are therefore interpreted as descriptive evidence about context–sensitive behavioural profiles rather than as reliability estimates in the conventional sense.

Exploratory factor analysis (EFA) was used to examine the preliminary internal structure of the indicators, serving an epistemological function: to interrogate whether the instrument's theoretically derived categories correspond to functional patterns in observed performance. Indicator counts were converted to within–session proportions, yielding 258 case–observations. Both varimax and oblimin rotations were performed (Costello & Osborne, 2005). Three methodological constraints should be noted. First, the Kaiser–Meyer–Olkin index (KMO = .596) is marginally below the conventional .60 threshold; the analysis is therefore exploratory rather than confirmatory. Second, case–observations are nested within a three–level structure (sessions within enrolments within 37 students, six of whom contributed to two modules), violating the EFA independence assumption at two levels. Third, within–session proportions are compositionally constrained, which can bias factor structure in known ways. These constraints define the EFA as construct

exploration, not validation (Borsboom et al., 2004; Cook et al., 2016; Kane, 2013). A secondary analysis excluding major contributions was also conducted. Given the explicitly exploratory purpose of the analysis and the theoretical interpretability of the emerging structure, both factors were retained for examination.

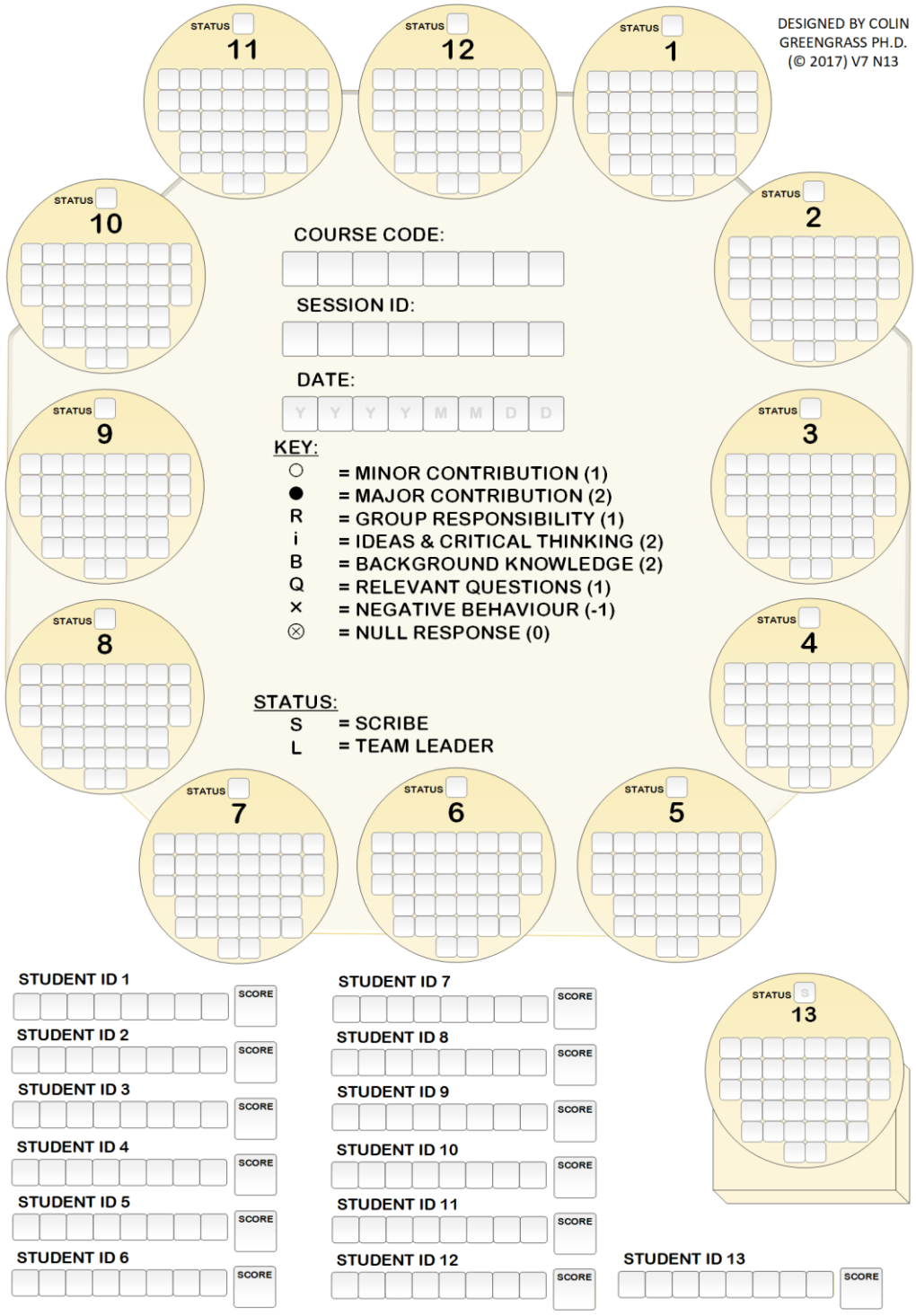

Figure 1. The EIPA-PBL recording schematic.

Positions represent students seated around the tutorial table, with the offset position representing the scribe. Within each student's cell, the tutor records contemporaneous events using the indicator key: minor contribution, major contribution, responsibility, ideas/critical thinking, background knowledge, relevant question, negative behaviour and null response. Numerical values indicate the low–stakes scoring weights used in the implementation reported here; the null marker is diagnostic and does not contribute to the score. The status field records formal tutorial role, such as scribe or leader, allowing indicator patterns to be interpreted alongside role assignment. The schematic is a recording schema, not a grading rubric: students are not expected to display every indicator within a single session. Detailed implementation guidance is available from the author.

Results

Pilot validation evidence

Two pilot cohorts of 10 students each were assessed prior to the main study to explore whether EIPA–PBL scores showed meaningful associations with external performance measures. It is of note that in the first cohort, EIPA–PBL total scores correlated more strongly with final examination performance ($r = .859$, $p < .001$) than did a comparator rubric–based assessment commonly used within the programme ($r = .636$, $p < .05$) – suggesting that the instrument may have captured performance information not represented by the rubric. In the second cohort, this was reinforced: EIPA–PBL scores correlated robustly with both seminar scores ($r = .886$, $p < .01$) and final examination scores ($r = .807$, $p < .05$), whilst the comparator rubric showed no significant correlation with either measure ($r = .18$ and $r = .177$ respectively). Within this second cohort, minor contributions ($r = .712$, $p < .05$) and background knowledge ($r = .759$, $p < .05$) correlated significantly with final examination scores, whilst ideas and questions showed positive but non–significant trends ($r = .626$ and $r = .643$ respectively), suggesting a theoretically interpretable pattern in which the more specific knowledge–oriented indicators were the stronger predictors. Given the small sample sizes, these findings are reported as preliminary construct–relevant evidence rather than validation in the psychometric sense.

Cross–session patterns

In the main study, intraclass correlation coefficients showed indicator–specific patterns across the six sessions. Minor contributions showed high cross–session repeatability (ICC = .884, 95% CI .82–.93). Background knowledge and questions showed moderate repeatability (ICC = .613, 95% CI .388–.778; ICC = .500, 95%

CI .193–.717). Ideas and creative thinking showed lower repeatability (ICC = .486, 95% CI .208–.698). Major contributions showed poor repeatability (ICC = .345, 95% CI –.005–.615), and responsibility showed very poor repeatability (ICC = .149, 95% CI –.370–.517).

These findings are consistent with the design expectation that role–dependent indicators would show lower cross–session stability than role–independent ones. Responsibility is expected to vary with formal role assignment: a student not assigned to a leader or scribe role may have fewer opportunities to display the behaviour in a given session. Major contributions are expected to vary with the case and the phase of the tutorial, as reporting opportunities cluster at particular moments and case complexity varies. Direct testing of these interpretations – by linking ICC values to per–session role assignment and case complexity – was not undertaken and represents a priority for further work. With this caveat, the ICC pattern is consistent with EIPA–PBL being sensitive to the shifting opportunities of tutorial taskwork rather than uniformly unreliable.

Consistency in scores for minor contributions demonstrated a tendency for students who made brief visible contributions in one session to do so similarly in others, suggesting a consistent tendency toward brief visible participation. The moderately good repeatability for background knowledge may more plausibly reflect recurring preparation or prior knowledge. The moderate repeatability for questions implies that subjects who ask questions will consistently do so across sessions, though as discussed below, this consistency may mask multidimensional constructs within the criterion.

Changes across consecutive sessions

A one–way repeated measures ANOVA across six consecutive sessions revealed significant differences for major contributions (Wilks' lambda = .427, $p < .001$, $\eta 2 = .156$), with session 5 showing notably higher counts compared to all other sessions – consistent with the interpretation that major contributions were frequently recording prepared reporting activity concentrated in specific phases. Background knowledge showed significant session effects, but these were not interpreted further because the study was not designed to separate session sequence from case content or tutorial phase. Ideas and creative thinking yielded a significant W–shaped cubic trend ($f(1,42) = 29.73$, $p < .001$, $\eta 2 = .157$), an interesting pattern which may reflect the alternating brainstorming and reporting structure of the sessions. Minor contributions showed no significant session differences. These patterns were calculated from data pooled across five cohorts. The present analysis cannot determine whether session differences reflected the content of particular cases, the roles students occupied,

or the position of the session within the module. Further work would need to model these effects separately.

Internal structure

Exploratory factor analysis loadings are reported in Table 1. Both varimax and oblimin solutions yielded a two–factor structure. Eigenvalues for the retained factors were 1.671 and 1.244, accounting for 34.50% and 24.89% of total variance respectively (approximately 59.4% combined).

Criterion	Varimax F1	Varimax F2	Oblimin F1	Oblimin F2	Oblimin (no major) F1	Oblimin (no major) F2
Minor contributions	.676	.174	.676	.167	.559	.361
Background knowledge	.752	–.170	.752	–.178	.724	.125
Ideas/creative thinking	.527	–.028	.527	–.034	.751	–.293
Questions	.409	.640	.408	.636	–.008	.904
Major contributions	–.322	.768	–.324	.772	–	–
KMO	.596		.596		.617	
Eigenvalues	1.671	1.244	1.671	1.244	1.724	.936
% total variance	34.50	24.89	34.50	24.89	43.11	23.40

Table 1. Exploratory factor analysis loadings for EIPA–PBL indicators. Loadings ≥ .30 are reported and interpreted as suggestive within an exploratory framework, in light of the methodological constraints described in Methods.

Three patterns in the loadings are notable. First, minor contributions, background knowledge and ideas and creative thinking cluster on the first factor. This is theoretically plausible: minor contributions – the low–inference real–time category – appear to be capturing brief instances of knowledge use or idea generation that were difficult to classify under more specific indicators during recording. Their clustering with knowledge and ideas suggests that the

pragmatic decision to include a low–inference category captured meaningful learning–relevant participation rather than random noise.

Second, major contributions loaded on the second factor and against the first (oblimin loadings of –.324 on factor 1 and .772 on factor 2). Rather than clustering with the knowledge–ideas group as a more substantial version of those behaviours, major contributions appears to represent a distinct construct. The component 1 axis is proposed to represent constructs based upon use and synthesis of background knowledge, whilst component 2 may represent a dimension in which higher–order collaborative cognition and lower–order reporting activity sit at opposing ends.

Third, questions loaded moderately on both factors (varimax .409/.640), suggesting it does not behave as a simple unitary construct. Removal of major contributions and re–running the oblimin analysis yielded a structure with KMO = .617, an improvement on the full–set value. In this reduced solution, ideas and creative thinking and background knowledge loaded strongly on the first factor whilst questions loaded almost exclusively on the second (oblimin .904).

Acceptability and implementation

In brief post-session focus-group discussions, students reported that EIPA-PBL provided timely feedback, allowed them to participate without consciously performing to assessment criteria, and encouraged greater engagement in the tutorial process. Open discussion also emphasised the importance of EIPA–PBL occupying a low–stakes position within the programme. Because the discussions were tutor–facilitated within an ongoing assessment relationship, they cannot be interpreted as independent evaluation; they are best treated as preliminary acceptability evidence.

EIPA–PBL was also trialled informally by two additional tutors, who reported that completing the schematic did not prevent monitoring discussion, interacting with students or redirecting thinking strategies (Hmelo-Silver & Barrows, 2006). Only the author's dataset was retained because no formal rater training or inter–rater reliability procedure had been established; inter–rater reliability therefore remains untested.

Discussion

Main contribution

This pilot study contributes an embedded real–time approach to a persistent problem in PBL tutorial assessment: how to generate individual–level evidence within a collaborative learning process. EIPA–PBL does not eliminate standardisation or judgement; it uses a shared indicator repertoire to record individual performance evidence as the tutorial unfolds, supporting profile–based assessment and targeted feedback. This framing preserves comparability whilst recognising that students may contribute differently to group task completion – addressing the long–recognised "individual within the collective" problem (Dijkstra et al., 2016; Strijbos, 2011) through shared construct, transparent indicators and individual evidence profiles.

Several existing instruments share related goals. TUTOTEST and shorter tutorial assessment tools provide structured post–session ratings of tutorial performance (Hébert & Bravo, 1996; Sim et al., 2006; St-Onge et al., 2014); CLeD–EX provides structured feedback on collaborative learning behaviours after the session (Pervaz Iqbal et al., 2020); and observational tools have been developed to code engagement through time–sampled snapshots of selected students (Alimoglu et al., 2014). To the author's knowledge, no published PBL assessment instrument has combined real–time event capture, schematic spatial recording of the seating arrangement, and indicator–type profiling within the live tutorial – these three design features combined within a shared framework to generate individual evidence profiles as the session unfolds.

Factor analysis as construct interrogation

Factor analysis is central to the present argument. It performs an epistemological function: interrogating whether the instrument's theoretically derived categories correspond to functional patterns in observed performance. As Borsboom et al. (2009) argued, many forms of assessment focus only upon adherence to nomological networks rather than interrogating whether the underlying behaviours actually correspond to the theorised constructs. Applied to EIPA–PBL, EFA was utilised to determine not what we believed was occurring within PBL sessions, but what is actually functioning as evidence within the data.

Where indicators cluster as expected, the structure is consistent with the intended interpretation. Where they cluster differently, the data are surfacing construct ambiguity, instrument refinement needs, or a theoretically interesting feature the original design did not anticipate; none of these outcomes is failure.

Validation must accumulate different types of evidence over multiple studies, and EFA contributes one source within this argument (Cook et al., 2016; Kane, 2013).

Indicator–by–indicator interpretation

The clustering of minor contributions with background knowledge and ideas and creative thinking is theoretically interpretable and, in this regard, represents a reassuring finding. The minor category was designed as a low–inference real–time backstop for relevant events that the tutor could not classify under a more specific indicator within the time available; its clustering with knowledge and ideas suggests that this design decision functioned precisely as intended, capturing learning–oriented acts rather than random noise.

The behaviour of major contributions is more challenging and theoretically informative. The indicator separates from the knowledge–ideas cluster and shows poor cross-session reliability — and it is of note that both findings converge on the same interpretation, observed by the author during scoring: that major contributions frequently captured extended reporting from prepared external materials during the reporting phase, rather than substantive participation in the collaborative reasoning process.

This raises a question that factor analysis can identify but not resolve: what is actually occurring when a student produces a major contribution? One possibility is that a student who contributed little to problem analysis may use the reporting phase to gain credit on the strength of reporting – a form of the silent free-riding problem (Czabanowska et al., 2012; Latané et al., 1979), in which strategic visibility substitutes for genuine collaborative cognition. A second possibility is that a student who understands the material but lacks confidence during analysis may hedge by reserving their contribution for the reporting phase, where the social risk is lower; marks are gained but preparation was genuine. The current instrument cannot distinguish among these, and the factor structure suggests it should not try to do so under a single criterion. A dedicated reporting criterion in future iterations would clarify whether reporting–phase behaviour correlates with, or substitutes for, collaborative–phase reasoning in the same student – and might address the more fundamental question of whether presentation of prepared external material constitutes valid evidence of collaborative cognition at all.

The *questions* criterion did not behave as a simple unitary construct, loading on both factors – consistent with the theoretical multidimensionality of questioning (Chin & Osborne, 2008; King, 1994; Schraw & Dennison, 1994), where some questions reflect high–level metacognitive monitoring, others are simple clarifications, and still others may be asked strategically to appear active

rather than from genuine inquiry. Tentatively, the second component may capture a cognitive–depth dimension along which clarification questions and extended reporting sit at opposing ends. Future development could split the criterion or retain it supplemented by anchor examples.

Responsibility showed very low cross–session repeatability, as expected when roles rotate. This supports interpreting the indicator as process regulation within specific task and role contexts rather than as a stable personal attribute. Future iterations should ensure that responsibility captures substantive actions that organise and guide the group's taskwork, rather than isolated visible gestures.

A validity argument for EIPA–PBL

Within Kane's argument–based validity framework (Kane, 2013; Messick, 1995), the present study contributes principally to three inferences. The scoring inference is supported in design: the structured indicator set, schematic format and operational definitions support reliable event recording, although same–event scoring agreement is not directly tested here. The generalisation inference is supported descriptively: multiple sessions can build a profile, although the number of sessions required for stability remains unknown. The extrapolation inference is partially supported: several indicators show internal–structure evidence consistent with intended interpretations; others (major contributions, questions) require refinement. The implications and consequences inferences are not yet addressed. Inter–rater reliability, response–process evidence and consequences research are required before stronger claims (Cook et al., 2016; Govaerts & van der Vleuten, 2013).

Performativity, equity and consequences

A live observational assessment may shape the behaviour it records. Most performativity research treats post–hoc rubric–driven assessment (Macfarlane, 2015; Torrance, 2007); if students treat the indicators as a checklist, they may speak strategically or interrupt inquiry to make performance visible. EIPA–PBL is therefore most defensibly used as a formative, portfolio or low–stakes summative tool within a broader programme (Schuwirth & van der Vleuten, 2011; Uijtdehaage & Schuwirth, 2018; van der Vleuten & Schuwirth, 2019).

The instrument attempts to avoid privileging talkativeness. Quiet students, those working in a second language, and those from cultures where rapid verbal assertion is less normative may contribute through synthesis, careful questioning or support for group process. Because EIPA–PBL is profile–based, such students can be visibly recorded under those indicators rather than penalised for low frequency on speech–dominant criteria. The schematic format

itself surfaces students with sparse records, providing a basis for tutor inquiry. The null marker distinguishes students who are repeatedly passed over from those who decline to respond when invited, prompting tutor questions about group facilitation strategy (Visschers-Pleijers et al., 2006).

Implications for use

EIPA–PBL suits feedback because it records specific behaviours rather than global impressions, and is available immediately after the session rather than reconstructed from memory. A student may be encouraged to move from brief factual contributions toward synthesis, to ask elaborative questions, or to take a more active role in responsibility for the group process. Across multiple sessions, EIPA–PBL generates a profile of how a student contributes across tasks, roles and cases – a basis for coaching aligning with sustainable assessment principles (Boud & Soler, 2016) and with the recommendation that tutors keep systematic records of a core set of behavioural domains (Eva et al., 2007). Low–stakes summative use may be defensible with repeated observations; high–stakes summative use is not currently defensible without further evidence.

Limitations and handover protocol

This study has several important limitations. The main analysed dataset came from a single tutor; inter–rater reliability was not established and is the most significant missing evidence for any stronger summative use. Student feedback was brief and tutor–facilitated and may have been shaped by the assessment relationship. EIPA–PBL records observable behaviours and cannot capture all learning, reflection or internal reasoning. The findings are based on a single PBL context. The sample comprised 37 unique students across 43 enrolment records; six students participated in two modules each, creating a three–level nesting structure that the EFA and ICC analyses were not designed to model explicitly. The EFA was additionally constrained by a sub–threshold KMO and compositional within–session proportions.

Although written consent was obtained and students retained the right to withdraw after the teaching relationship had ended, the study was conducted in a setting where teaching, assessment and research roles overlapped. This may have influenced perceptions of voluntariness despite the zero–consequence design. Future studies should consider independent consent procedures and independently facilitated feedback.

The author no longer works in a PBL programme and cannot extend data collection in this setting; the present article therefore positions EIPA–PBL as a handover framework for further evaluation by other PBL programmes. A

minimal next–stage programme should include: (1) rater training using written vignettes or video clips; (2) same–event scoring studies with multiple raters; (3) repeated live tutorial sampling across cases and roles; and (4) consequences research examining performativity, perceived fairness and group dynamics – steps that map directly onto the inferences of an argument–based validity programme.

Conclusion

EIPA-PBL provides a candidate approach for recording individual performance within live PBL tutorial activity. Its contribution lies in combining real-time event capture, schematic spatial recording and indicator-type profiling within a shared framework for individualised assessment. The pilot findings suggest that several indicators functioned in theoretically interpretable ways, while others, particularly major contributions and questions, may require refinement. Factor analysis is therefore used not to confirm the instrument as validated, but to examine how theoretically derived categories functioned as evidence within observed tutorial performance. In this respect, EIPA-PBL offers a way of testing assumptions about what appears to matter in PBL tutorials against the patterns generated by actual recorded events. At this stage, EIPA-PBL should be understood as a formative, portfolio and bounded low-stakes assessment tool. Further work should establish inter-rater reliability, response-process evidence, consequences for group dynamics and the conditions under which profile-based evidence can support defensible judgement across PBL contexts.

References

Alimoglu, M. K., Sarac, D. B., Alparslan, D., Akman Karakas, A., & Altintas, L. (2014). An observation tool for instructor and student behaviors to measure in-class learner engagement: A validation study. *Medical Education Online*, *19*, 24037. https://doi.org/10.3402/meo.v19.24037

Barrows, H. S., & Tamblyn, R. M. (1980). *Problem-based learning: An approach to medical education*. Springer.

Bearman, M., & Ajjawi, R. (2018). From seeing through to seeing with: Assessment criteria and the myths of transparency. *Frontiers in Education*, *3*, 96. https://doi.org/10.3389/feduc.2018.00096

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 135–170). Information Age Publishing Inc. https://doi.org/10.1108/978-1-61735-269-020251010

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295x.111.4.1061

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, *41*(3), 400–413. https://doi.org/10.1080/02602938.2014.999746

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Chin, C., & Osborne, J. (2008). Students' questions: A potential resource for teaching and learning science. *Studies in Science Education*, *44*(1), 1–39. https://doi.org/10.1080/03057260701828101

Cook, D. A., Kuper, A., Hatala, R., & Ginsburg, S. (2016). When assessment data are words: Validity evidence for qualitative educational assessments. *Academic Medicine*, *91*(10), 1359–1369. https://doi.org/10.1097/acm.0000000000001175

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(1), 7. https://doi.org/10.7275/jyj1-4868

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Czabanowska, K., Moust, J. H., Meijer, A. W., Schröder-Bäck, P., & Roebertsen, H. (2012). Problem-based learning revisited: Introduction of active and self-directed learning to reduce fatigue among students. *Journal of University Teaching & Learning Practice*, *9*(1), Article 6. https://doi.org/10.53761/1.9.1.6

de Hei, M., Strijbos, J.-W., Sjoer, E., & Admiraal, W. (2016). Thematic review of approaches to design group learning activities in higher education: The development of a comprehensive framework. *Educational Research Review*, *18*, 33–45. https://doi.org/10.1016/j.edurev.2016.01.001

Dijkstra, J., Latijnhouwers, M., Norbart, A., & Tio, R. A. (2016). Assessing the 'I' in group work assessment: State of the art and recommendations for practice. *Medical Teacher*, *38*(7), 675–682. https://doi.org/10.3109/0142159x.2016.1170796

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 1–19). Elsevier.

Dolmans, D. H. J. M., De Grave, W., Wolfhagen, I. H. A. P., & van der Vleuten, C. P. M. (2005). Problem-based learning: Future challenges for educational practice and research. *Medical Education*, *39*(7), 732–741. https://doi.org/10.1111/j.1365-2929.2005.02205.x

Eva, K. W. (2001). Assessing tutorial-based assessment. *Advances in Health Sciences Education*, *6*(3), 243–257. https://doi.org/10.1023/A:1012743830638

Eva, K. W., Solomon, P., Neville, A. J., Ladouceur, M., Kaufman, K., Walsh, A., & Norman, G. R. (2007). Using a sampling strategy to address psychometric challenges in tutorial-based assessments. *Advances in Health Sciences Education*, *12*(1), 19–33. https://doi.org/10.1007/s10459-005-2327-z

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: Assessor cognition from three research perspectives. *Medical Education*, *48*(11), 1055–1068. https://doi.org/10.1111/medu.12546

Govaerts, M. J. B., & van der Vleuten, C. P. M. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education*, *47*(12), 1164–1174. https://doi.org/10.1111/medu.12289

Hébert, R., & Bravo, G. (1996). Development and validation of an evaluation instrument for medical students in tutorials. *Academic Medicine*, *71*(5), 488–494. https://doi.org/10.1097/00001888-199605000-00020

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, *16*(3), 235–266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

Hmelo-Silver, C. E., & Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. *Interdisciplinary Journal of Problem-Based Learning*, *1*(1), 21–39. https://doi.org/10.7771/1541-5015.1004

Hmelo-Silver, C. E., & Barrows, H. S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction*, *26*(1), 48–94. https://doi.org/10.1080/07370000701798495

Hutchins, E. (1995). *Cognition in the wild*. MIT Press. https://doi.org/10.7551/mitpress/1881.001.0001

Johnson, D. W., & Johnson, R. T. (2005). New developments in social interdependence theory. *Genetic, Social, and General Psychology Monographs*, *131*(4), 285–358. https://doi.org/10.3200/MONO.131.4.285-358

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706. https://doi.org/10.1037/0022-3514.65.4.681

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American*

Educational Research Journal, *31*(2), 338–368. https://doi.org/10.3102/00028312031002338

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, *7*(3), 77–124. https://doi.org/10.1111/j.1529-1006.2006.00030.x

Kulhavy, R. W., & Stock, W. A. (1996). How cognitive maps are learned and remembered. *Annals of the Association of American Geographers*, *86*(1), 123–145. https://doi.org/10.1111/j.1467-8306.1996.tb01748.x

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*(6), 822–832. https://doi.org/10.1037/0022-3514.37.6.822

MacDonald, R. (2005). Assessment strategies for enquiry and problem-based learning. In T. Barrett, I. Mac Labhrainn, & H. Fallon (Eds.), *Handbook of enquiry and problem-based learning: Irish case studies and international perspectives* (pp. 85–94). Centre for Excellence in Learning and Teaching, NUI Galway, in association with the All Ireland Society for Higher Education.

Macfarlane, B. (2015). Student performativity in higher education: Converting learning as a private space into a public performance. *Higher Education Research & Development*, *34*(2), 338–350. https://doi.org/10.1080/07294360.2014.956697

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356–376. https://doi.org/10.5465/amr.2001.4845785

Mercer, N. (2000). *Words and minds: How we use language to think together*. Routledge.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nendaz, M. R., & Tekian, A. (1999). Assessment in Problem-Based Learning Medical Schools: A Literature Review. *Teaching and Learning in Medicine*, *11*(4), 232–243. https://doi.org/10.1207/S15328015TLM110408

Nijstad, B. A., & Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review*, *10*(3), 186–213. https://doi.org/10.1207/s15327957pspr1003_1

Norman, G., Bordage, G., Page, G., & Keane, D. (2006). How specific is case specificity? *Medical Education*, *40*(7), 618–623. https://doi.org/10.1111/j.1365-2929.2006.02511.x

Pervaz Iqbal, M., Velan, G. M., O'Sullivan, A. J., & Balasooriya, C. (2020). The collaborative learning development exercise (CLeD-EX): An educational

instrument to promote key collaborative learning behaviours in medical students. *BMC Medical Education*, *20*(1), 62. https://doi.org/10.1186/s12909-020-1977-0

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448. https://doi.org/10.1080/02602930902862859

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a big five in teamwork? *Small Group Research*, *36*(5), 555–599. https://doi.org/10.1177/1046496405277134

Savery, J. R. (2006). Overview of problem-based learning: Definitions and distinctions. *Interdisciplinary Journal of Problem-Based Learning*, *1*(1), 9–20. https://doi.org/10.7771/1541-5015.1002

Savin-Baden, M. (2004). Understanding the impact of assessment on students in problem-based learning. *Innovations in Education and Teaching International*, *41*(2), 221–233. https://doi.org/10.1080/1470329042000208729

Schmidt, H. G. (1983). Problem-based learning: Rationale and description. *Medical Education*, *17*(1), 11–16. https://doi.org/10.1111/j.1365-2923.1983.tb01086.x

Schmidt, H. G., Rotgans, J. I., & Yew, E. H. J. (2011). The process of problem-based learning: What works and why. *Medical Education*, *45*(8), 792–806. https://doi.org/10.1111/j.1365-2923.2011.04035.x

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, *33*(6), 478–485. https://doi.org/10.3109/0142159X.2011.565828

Scott, B. M., & Schwartz, N. H. (2007). Navigational spatial displays: The role of metacognition as cognitive load. *Learning and Instruction*, *17*(1), 89–105. https://doi.org/10.1016/j.learninstruc.2006.11.008

Segers, M., & Dochy, F. (2001). New Assessment Forms in Problem-based Learning: The value-added of the students' perspective. *Studies in Higher Education*, *26*(3), 327–343. https://doi.org/10.1080/03075070120076291

Sim, S. M., Azila, N. M. A., & Lee, M. S. (2006). A simple instrument for the assessment of student performance in problem-based learning tutorials. *Annals of the Academy of Medicine, Singapore*, *35*(9), 634–641. https://doi.org/10.47102/annals-acadmedsg.V35N9p634

Slavin, R. E. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology*, *21*(1), 43–69. https://doi.org/10.1006/ceps.1996.0004

St-Onge, C., Frenette, E., Cote, D. J., & De Champlain, A. (2014). Multiple tutorial-based assessments: A generalizability study. *BMC Medical Education*, *14*, 30. https://doi.org/10.1186/1472-6920-14-30

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. MIT Press. https://doi.org/10.7551/mitpress/3372.001.0001

Stewart, D. W., Shamdasani, P. N., & Rook, D. W. (2007). *Focus groups: Theory and practice, 2nd ed*Sage Publications, Inc. https://doi.org/10.4135/9781412991841

Strijbos, J.-W. (2011). Assessment of (computer-supported) collaborative learning. *IEEE Transactions on Learning Technologies*, *4*(1), 59–73. https://doi.org/10.1109/tlt.2010.37

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education*, *18*(2), 291–303. https://doi.org/10.1007/s10459-012-9370-3

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice*, *14*(3), 281–294. https://doi.org/10.1080/09695940701591867

Uijtdehaage, S., & Schuwirth, L. W. T. (2018). Assuring the quality of programmatic assessment: Moving beyond psychometrics. *Perspectives on Medical Education*, *7*(6), 350–351. https://doi.org/10.1007/s40037-018-0485-y

Van Berkel, H. J. M., & Schmidt, H. G. (2000). Motivation to commit oneself as a determinant of achievement in problem-based learning. *Higher Education*, *40*(2), 231–242. https://doi.org/10.1023/A:1004022116365

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2019). Assessment in the context of problem-based learning. *Advances in Health Sciences Education*, *24*(5), 903–914. https://doi.org/10.1007/s10459-019-09909-1

van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education: Design principles and strategies. *Medical Education*, *44*(1), 85–93. https://doi.org/10.1111/j.1365-2923.2009.03498.x

Visschers-Pleijers, A. J. S. F., Dolmans, D. H. J. M., De Grave, W. S., Wolfhagen, I. H. A. P., Jacobs, J. A., & van der Vleuten, C. P. M. (2006). Student perceptions about the characteristics of an effective discussion during the reporting phase in problem-based learning. *Medical Education*, *40*(9), 924–931. https://doi.org/10.1111/j.1365-2929.2006.02548.x

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research*, *13*(1), 21–39. https://doi.org/10.1016/0883-0355(89)90014-1