3.5

Observations and the Assessment of Behaviours

Francesco Ciriello[a], Ali Jaber[b], and Claire Lucas[c]

[a] King's College London, United Kingdom, ORCID: 0000-0003-1288-3114;
[b] Aston University, United Kingdom;
[c] Aston University, United Kingdom, ORCID: 0000-0003-2284-272X.

Keywords: *observational assessment, behavioural marker systems, assessment methods, competency evaluation, behavioural assessment, transferable competencies and skills*

Observations involve the deliberate, systematic, and purposeful act of watching and listening to learners as they engage in tasks relevant to their work, enabling educators to assess and provide feedback on learners' competencies and to track how these competencies develop over time. This chapter is written for engineering educators and assessors who wish to incorporate observational assessment into their practice, with particular emphasis on non-technical skills. It explores how behaviours are observed within educational settings, with particular attention to Behavioural Marker Systems (BMS) as structured observational frameworks for competencies such as communication, teamwork, leadership, transferable competencies and skills, and problem solving. The chapter guides readers through principles, methods, and pragmatic steps for planning and conducting observations of non-technical skills, and for establishing and validating a BMS for engineering contexts.

DOI: 10.59490/mt.266.80.

1 Introduction

Engineering education is fundamentally a social process. It unfolds through a series of sociotechnical interactions, where learners engage with instructors, peers, learning materials, and their environment. These interactions vary in their degree of planning and purpose, and as an educator, you have likely experienced how profoundly they are shaped by learners' behaviours.

You invest significant effort in creating constructive environments to support learning. But have you also built the right infrastructure to observe your learners within these settings? From the front of a lecture theatre, you might see students scribbling notes, yawning, or chatting with peers. But how can you tell if they are learning?

Observations involve the deliberate, systematic, and purposeful act of watching and listening to learners as they engage in tasks relevant to their work. Despite their value, observations are often overlooked by educators, who may feel uncertain about how to select appropriate observational frameworks or implement them effectively in their own context. The purpose of this chapter is to demystify observational assessment and provide practical guidance on how to embed it into your teaching practice.

In this chapter, we help you formalise the way you plan and conduct observations of learning, particularly in the context of assessing competencies. We remain flexible in our approach, highlighting guiding principles for good practice and grounding these techniques in educational research. Specifically, we begin in Section 2 by defining observational methods and outlining the hierarchy of approaches available to educators. Section 3 explores how to establish robust observational frameworks, drawing on research methodologies while acknowledging their limitations and nuances. We turn to a practical case study of a Behavioural Marker System (hereafter, BMS) in development from our research at King's College London, before broadening our discussion to offer general recommendations for implementing good practice. While much of the BMS literature originates from industry contexts (and indeed, you will notice that we have drawn from several industrial examples in our references), we believe their principles apply equally well to educational settings and will help you understand how an engineering observational framework was created within a Higher Education context.

This chapter is written for engineering educators and assessors who are seeking to incorporate observational assessment into their programmes, whether in project-based modules, laboratories, design studios, or workplace learning. Observational assessment is particularly valuable for competencies highlighted in the SEFI handbook and the Engineering Education Research (EER) taxonomy, such as communication, teamwork, leadership, and professional skills, because these outcomes are enacted through observable behaviour in context rather than solely through written products or tests. For problem solving, for instance, observation allows assessors to see how students define problems, seek information, and respond to uncertainty during authentic tasks, complementing product-focused assessments.

2 Definition of observational assessment

Observational assessment is widely recognised as a fundamental method for evaluating human performance, particularly in high-stakes, high-reliability domains such as aviation and healthcare (Flin, O'Connor & Chrichton, 2017). It offers a structured approach to observe, record, interpret, and evaluate an individual's or team's actions against predefined standards. In this chapter, we interpret behaviour as the way individuals enact a competency through their skills, attitudes, and knowledge. In Chapter 1.2 of the handbook, competencies and their many definitions have been discussed at length. One such definition states that while competencies are understood to encompass knowledge, skills, and attitudes (Crichton et al., 2017), attitudes and knowledge are internal states that cannot be directly observed. Observation provides the critical link between these internal competencies and the external behaviours that demonstrate them. Behavioural assessment therefore aims to infer these internal competencies from external cues — what individuals say, do, or omit — while performing a task.

By systematically studying human interactions with tasks, equipment, systems, and one another, observation captures evidence of competency in action during training, licensing, and ongoing professional development. This process typically involves watching individuals or teams as they perform tasks in either real-world or simulated environments, enabling the collection of rich data on behaviours, decision making, and collaboration. Observational methods can be qualitative or quantitative, self-directed or externally conducted, depending on the purpose and level of detail required (Flin et al., 2017).

Building on this premise, we will guide you through how we have worked on Behavioural Marker Systems , as they offer structured taxonomies of skills, elements, and associated behaviours for training and assessment (Flin et al., 2006). They are grounded in the idea that observable behaviours provide evidence of underlying skills and knowledge (Tefera, 2019) and are thus particularly interesting for practitioners who want to understand the formalisms of observational assessment science.

A BMS is a structured framework comprising categories of non-technical skills, their behavioural elements, and associated poor and good performance indicators, designed to support systematic observation, assessment, and formative feedback in authentic tasks (Hamlet et al., 2023).

BMS are introduced here as a case study because they exemplify how observational assessment can be systematically designed to capture complex competencies, bridge the gap between observable actions and underlying skills, and provide actionable insights for both training and performance evaluation. Their structured approach allows for practical demonstration of the principles discussed in this chapter, making abstract concepts of competency assessment tangible and operational. Before introducing how to work with BMS further (Section 3), let us discuss the context and limitations of observational assessments at a general level.

2.1 Using observations for assessment

The observation of behaviour as evidence of competency can be approached using different techniques. At one end of the spectrum are structured methods, such as predefined checklists designed to help observers identify specific cues. At the other end are open-ended approaches, where observers record free-text notes based on their observations. A mixed approach can also be employed, combining the strengths of both methods. Common data collection techniques include anecdotal records, rating scales, checklists, event and time sampling, as well as the use of live, recorded, or instrumented (technology-enabled) observations. The role of the observer may also vary, ranging from participatory, where the observer is involved in the activity being studied, to external, where they remain detached from the process.

Specialised observation formats have been developed to enable the systematic use of behaviours for assessment and feedback. One such approach is the use of behavioural marker systems, structured observational tools originating in medical and emergency training contexts. Behavioural marker systems provide structured frameworks for observing and rating specific actions or behaviours. By organizing these behaviours into skill categories, elements, and observable performance indicators, they support consistent training, assessment, and feedback (Fletcher et al., 2003).

Beyond BMS, several other approaches exist for assessing and developing competencies. Assessment centres, for example, are structured processes in which individuals complete multiple tasks, such as role-plays, group discussions, and interviews, designed to evaluate specific competencies. Trained assessors observe and score behaviours to provide a reliable and well-rounded picture of an individual's suitability for a job, promotion, or development opportunity (Ballantyne & Povah, 2017). Development centres share many similarities with assessment centres but focus more explicitly on individual growth and learning rather than selection or recruitment. Through exercises such as role-plays, feedback sessions, and psychometric tests, these centres help identify strengths and areas for development, offering participants targeted support for improvement.

Finally, the increasing use of digital technologies has enabled the emergence of instrumented teamwork approaches, such as those explored in the MIT System Design & Management programme (Graham 2022). Here, digital tools and data collection systems are employed to monitor, support, and assess collaborative learning activities. By capturing interactions such as communication patterns, task contributions, and coordination efforts, these tools provide students and educators with feedback that can be used to improve teamwork processes and enable evidence-based evaluation.

Table 3.5:1: Commonly used observational methods discussed in this section, with indicative uses and limitations in engineering education

Method	*Description*	*Suggested use in engineering education*	*Limitations*
Structured checklists	Predefined behaviours list; tick-off during observation (Flin et al., 2017)	Safety monitoring in labs; coverage in design-build-test tasks (Flin et al., 2017)	Superficial box-ticking risk; misses quality and context adaptation (Flin et al., 2017)
Open-ended field notes	Free-text behaviour descriptions; written during or after observation	Early BMS development; piloting project-based modules	Time-consuming to analyse; inconsistent across observers
Rating scales	Numerical/descriptive scales; anchored with performance descriptors (Hamlet et al., 2023)	Teamwork, communication, leadership; group projects and design reviews	Needs clear anchors and rater training; reliability needs calibration (Hamlet et al., 2023)
Anecdotal records	Brief, specific incident accounts; illustrates key competencies	Conflict management moments; ethical decision points in projects	Underrepresents routine positives; subject to selective recall
Event or time sampling	Records at set time intervals or when specific events occur	Tracking feedback-seeking frequency; peer involvement in lab sessions (Flin et al., 2017)	Increases observer workload; may miss infrequent behaviours
Video-based observation	Recorded/live-capture scenarios; supports repeated reflective review (Hamlet et al., 2023)	Simulation debriefings; capstone reviews and assessor calibration	Camera/audio may obscure cues; over-focus on visible behaviours
Behavioural Marker Systems (BMS)	Taxonomy of non-technical skills; behavioural elements and rating scales (Hamlet et al., 2023; Tefera, 2019)	Communication, teamwork, ethics; design projects, labs, simulations	High development effort; high cognitive load for new assessors

2.2 Observability and its limitations

Observational assessment is a powerful tool for evaluating human performance, but it has recognised limitations. A central challenge is the invisibility of covert cognitive processes: decision making, planning, and other internal mental activities are not directly observable. While these processes can sometimes be inferred from communication or external actions, they remain largely hidden to observers (Thompson, 2013).

Furthermore, the capabilities of human observers impose practical constraints. Observers can experience cognitive overload, fatigue, distraction, or unconscious bias, particularly in high-pressure or complex environments (O'Connor et al., 2002; Thompson, 2013). These factors can compromise the accuracy, consistency, and fairness of assessments. Subtle or infrequent behaviours can be missed altogether, while very frequent behaviours may need capping to prevent excessive recording and maintain focus. Detecting omissions, i.e., the absence of expected action, requires especially skilled observers familiar with the task. As Thompson (2013, p. 113) notes: "*Through focusing specifically on the recording of overt explicit behaviour, the method precludes the direct recording of omissions… The detection of omitted behaviour lends itself more to scripted test environments where the occurrence (or not) may be anticipated by the observer.*"

These challenges are further amplified in virtual environments. Physical, cognitive, and psychological 'virtualisation' can reduce social presence, obscure non-verbal cues, and fragment attention, making it harder to interpret collaborative dynamics or engagement authentically. Observers must adapt frameworks and expectations to ensure fair and meaningful assessments in these contexts.

An important design principle in observational assessment is the focus on behavioural patterns rather than personality traits or attitudes. Behavioural marker systems (BMS) and similar tools should describe specific, observable behaviours that reflect skills or knowledge in action, not subjective traits (Klampfer et al., 2001). While personality influences behaviour, assessment and training interventions target modifiable behavioural patterns that impact performance rather than attempting to change inherent personality characteristics (Flin et al., 2017). As Flin et al. (2017, p. 9) note: "*The interventions… are about changing patterns of behaviour rather than attempting to modify the individual's underlying personality.*"

For example, an autocratic leader may need to adopt behaviours that encourage team input, while introverts may need to practise outward communication, and extroverts may need to focus on active listening (Flin et al., 2017). Distinguishing between personality and behaviour ensures objectivity and keeps assessments development-oriented rather than judgemental.

The validity and reliability of observational assessments depend heavily on how observations are structured and interpreted. Bias can emerge through halo effects, confirmation bias, or cultural and gender stereotypes, especially in unstructured settings or peer assessments. Structured rubrics, observer training, multiple assessors, and reflective practices help mitigate these risks, particularly in complex domains like teamwork and professionalism (O'Connor & Long, 2011).

The modality of observation also shapes what cues are available. In-person observations capture full non-verbal communication, including body language and eye contact, while video may omit peripheral or spatial cues. Audio-only modalities, in turn, rely entirely on tone, pace, and language, excluding visual signals. Because non-verbal cues are context-dependent and culturally variable, observers must consider task context, relationships, and cultural norms to avoid misinterpretation.

To enhance validity and insight, educators can employ observational lenses, i.e., structured perspectives for analysing practice. Aligned with Brookfield's (1995) four lenses of feedback — autobiographical, student, colleague, and theoretical — these lenses help educators interpret complex interactions from multiple viewpoints, making observations more focused, actionable, and reflective.

Observational frameworks are most powerful for formative assessment, feedback, and professional development, where the goal is reflection and improvement rather than high-stakes judgement. Their structured nature promotes transparency, alignment with learning outcomes, and consistency across multiple assessors. However, the subjectivity of observation and the influence of bias and context mean that for summative assessment, robust safeguards, such as clear criteria, rater training, and moderation, are essential to ensure reliability. With these safeguards, observational frameworks can meaningfully contribute to both formative and summative assessment, especially for complex competencies such as teamwork, communication, and professionalism.

Observations can be longitudinal, capturing behavioural patterns and growth over time, offering insights into development and consistency but requiring sustained effort and observer continuity. They can also be event-based, focusing on specific moments like presentations or simulations, offering efficiency and detail but risking misrepresentation if the event is uncharacteristic. Both approaches are complementary, and their integration provides a fuller understanding of performance.

Finally, observational assessments may target individual performance, team performance, or both. For instance, systems like "*Non-Technical Skills*" (NOTECHS) assess individual pilots (Flin et al., 2003), while the Anaesthetist's Non-Technical Skills (ANTS) focuses on individual anaesthetists (Fletcher et al., 2004). In contrast, many BMS frameworks, such as those developed for drilling teams (Crichton et al., 2017) or search-and-rescue crews (Hamlet et al., 2023), emphasise collective competencies such as coordination, communication, and adaptability, which are essential for achieving shared goals.

3 GENTS-BMS: steps towards developing an observational framework

Behavioural Marker Systems are well established in professional and safety-critical education, including healthcare, aviation, emergency services, and other domains where non-technical skills are learned through practice. Their use in engineering higher education is much less developed. In this section, we will consider the example of the *General Engineering Non-Technical Skills Behavioural Marker System (GENTS-BMS)* that we developed during PhD research at King's College London between 2022 and 2026 to implement in the General Engineering undergraduate programme of our engineering department. We hope this demonstrates the different ways in which you can establish and validate an observational framework and can support your development of a research-informed framework to use in teaching.

The example therefore does not claim to provide a fully validated engineering education instrument for you to readily use in your teaching, albeit you can find example implementations in the reference videos of Lucas & Ciriello (2026). Instead, it illustrates how an engineering educator might translate BMS principles into a usable observational framework for project-based, laboratory, studio, or simulation-based engineering learning.

3.1 Context of the case study

Our *GENTS-BMS* was developed through an iterative process that combined literature, expert knowledge, and practitioner feedback (Lucas & Ciriello, 2026). We started by reviewing non-technical skills to from an initial set of categories and elements, which we then presented in semi-structured interviews to senior engineers and engineering leaders across the UK to elicit concrete behavioural examples of good and poor performance, each mapped onto skill categories. We also ran validation workshops at engineering education conferences (EERN, SEFI), which involved educators sorting and critiquing behavioural elements, refining wording, and checking recognisability and distinctiveness of categories. Curriculum integration is still ongoing, and further work is underway to evaluate its application in educational settings, refine the behavioural markers based on implementation feedback, and establish how the framework can best support the development and assessment of non-technical skills in engineering students.

3.2 Methodologies to establish a Behavioural Marker System

Establishing a BMS is a rigorous process that typically involves several integrated methodologies to ensure the system is comprehensive, context-specific, and accurately reflects real-world performance (Hamlet et al., 2023). As an educator, you will be confronted with several choices for how to establish your observational framework, and indeed you should seek to draw from good practices across a range of steps, particularly in how you draw behaviours from literature, events, tasks, interviews, focus groups or direct observation.

The aim for a BMS is to define observable, non-technical behaviours that contribute to superior or substandard performance within a specific work environment (Dias et al., 2023), and there are many ways to approach these definitions.

Literature review

As an initial step, you will survey existing scientific literature to identify established non-technical skills taxonomies and behavioural marker systems developed in similar or related domains. This helps in establishing a baseline understanding of common categories and elements of behaviour already recognised in the field (Yule et al., 2006).

To establish our GENTS-BMS skills taxonomy, we conducted a systematic literature review of non-technical skills literature in engineering education, and used frequency of mentions to establish skills categories and elements (Jaberi, Ciriello & Lucas, 2024). When approaching the literature review, you will quickly notice that the field lacks conceptual clarity and you might be rightfully be worried that any meaningful assessment is only possible when concepts are clearly defined. For example, 53% of papers we reviewed did not explicitly define the soft skills they investigated.

The process of building a taxonomy is therefore as much about clarifying concepts and boundaries as it is about identifying skills. A useful taxonomy distinguishes between underlying dispositions and observable behaviours, is grounded in disciplinary context, and is designed from the outset to support assessment and development. The move from broad skill labels to observable behaviours is where a taxonomy becomes usable.

Event-based analysis (incidents and accidents reports analysis)

In a high-stake environment, event-based analysis, specifically the examination of incidents and accident reports, constitutes a methodologically rigorous approach for identifying behavioural patterns that either contribute to, or mitigate against, adverse outcomes, human error, and safety-related events (O'Connor & Long, 2011, Tefera, 2019). This approach entails the systematic collection of documented analyses from organisations, alongside incident data and formal reports (Thompson, 2013).

Within these reports, human factors are scrutinised with a focus on isolating procedural lapses, errors, or deviations that can be attributed to identifiable behavioural antecedents. The analytic process further involves the extraction of particular instances in which non-technical skills such as communication, decision making, or situational awareness either exacerbated the incident or facilitated its mitigation (Hamlet et al. 2023).

We think it is worth acknowledging that this is often the ideal method, but one that is not always feasible, particularly outside high-risk industries where access to incident data may be restricted or where formal reporting systems are less mature (indeed we did not use this event-based analysis in GENT-BMS). Access to incident reports is often restricted due to confidentiality, and the systematic coding and analysis of large numbers of reports requires considerable time and expertise. Practitioners developing a BMS may therefore adopt a more pragmatic approach by reviewing publicly available accident investigation reports, published case studies, regulatory safety bulletins, or organisational lessons-learned documents. Although these sources may not provide the same breadth or contextual richness as internal incident databases, they can still yield valuable insights into the behaviours associated with successful and unsuccessful performance and help validate or refine an emerging taxonomy.

Cognitive Task Analysis (CTA)

One of the most effective ways to develop behavioural markers is to talk directly to experienced practitioners. Many aspects of non-technical performance are tacit: experts recognise good communication, sound judgement or effective leadership when they see it, but rarely describe these behaviours explicitly. Cognitive Task Analysis (CTA) provides a structured way of making this expertise visible by encouraging experts to reflect on how they think, make decisions and behave in real situations (Yule et al., 2006).

Rather than asking experts to describe what makes a 'good engineer', ask them to recall specific incidents they have experienced. While confidentiality constraints may prevent practitioners from sharing details of specific workplace incidents, they are often able to discuss these experiences in more general terms by describing representative or idealised scenarios. Semi-structured interviews, combined with the Critical Incident Technique (CIT), are particularly well suited to this purpose because they encourage discussion of real examples rather than abstract opinions. Ask participants to describe both successful and unsuccessful situations, then probe how they recognised the problem, what information they considered, what actions they took, why they chose those actions, and how the outcome might have been different. This approach helps uncover the behaviours that distinguish stronger and weaker performance (Hamlet et al., 2023).

In developing GENTS-BMS, we interviewed experienced engineers and engineering educators from industry, higher education, recruitment, and professional accreditation, ensuring that each non-technical skill domain was discussed by individuals with substantial expertise. Rather than beginning with a blank sheet of paper, interviews were structured around the emerging non-technical skills taxonomy developed through the literature review and expert workshops. Each participant was first asked to describe the characteristics of a high-performing engineer before discussing one specific non-technical skill category in depth. For every skill element, participants were invited to describe how high-performing and lower-performing engineers behave in practice, allowing positive and negative behavioural markers to be elicited in parallel.

A key lesson from this work was the importance of continually redirecting participants towards observable behaviour. Experts naturally tend to describe attributes such as 'good communicator' or 'strong leader'. Your role as interviewer is to probe further: What did they actually do? What questions did they ask? What did they say in the meeting? What would I have observed if I had been standing next to them? These follow-up questions transform abstract competencies into behavioural markers that can later be observed and assessed. This behaviour-focused questioning was central to the development of GENTS-BMS and is consistent with knowledge elicitation approaches used in behavioural marker systems developed for aviation and healthcare.

Finally, record and transcribe the interviews before systematically coding the data against your developing taxonomy. We recommend distinguishing examples of effective and ineffective performance within each skill element, then rewriting these as short, observable behavioural statements. Avoid statements that describe personality or internal thought processes; instead, write markers that describe actions someone could realistically observe during a project, laboratory session or design review. These behavioural statements become the foundation of your behavioural marker system.

Focus groups

Once you have developed an initial set of non-technical skill categories and behavioural markers, the next step is to test them with a wider group of experts. Focus groups and facilitated workshops provide an opportunity to validate your emerging framework, challenge assumptions, and build consensus about whether the behaviours accurately reflect effective professional practice (Butler et al., 2020).

Rather than presenting participants with a blank sheet of paper, provide them with the draft taxonomy and behavioural markers generated through your literature review, interviews, and other evidence sources. This gives the discussion a clear focus and allows participants to concentrate on refining the framework rather than creating it from scratch. Encourage participants to question whether each behavioural marker is clear, observable, and representative of the skill it is intended to describe. Ask whether important behaviours are missing, whether some markers overlap, or whether behaviours would be better assigned to another skill category (Hamlet et al., 2023; Irwin et al., 2023).

In developing GENTS-BMS, we used workshops with engineering educators, industrial partners, professional body representatives, and other engineering stakeholders to iteratively refine the framework. Participants worked through each non-technical skill category, discussing the proposed behavioural markers and suggesting additions, modifications, and rewording where appropriate. These discussions also helped identify behaviours that appeared under multiple skills, allowing the framework to be simplified and improving the distinction between categories. Repeated review across multiple workshops gradually strengthened both the content and organisation of the behavioural marker system.

A key advantage of group discussion is that it exposes differences in perspective. Industry practitioners, educators, employers, and professional bodies often emphasise different aspects of engineering practice. Rather than viewing these differences as disagreement, treat them as an opportunity to develop behavioural markers that are applicable across a range of engineering contexts. Where consensus cannot immediately be reached, record the differing viewpoints and revisit them during subsequent iterations rather than forcing agreement during the workshop.

As the discussion progresses, continually return participants to observable behaviours. Experts frequently propose statements such as 'demonstrates professionalism' or 'shows leadership'. Encourage them to explain how this would be observed in practice. Ask questions such as: How would you recognise this behaviour? What would you expect to see or hear? What would distinguish excellent from poor performance? These prompts help transform broad competencies into behavioural markers that can be observed consistently by different assessors.

Finally, document all proposed changes and revise the framework after each workshop. It is often more effective to undertake several rounds of focused review with smaller groups than to attempt to produce a final framework in a single session. An iterative process of discussion, revision, and revalidation helps ensure that your behavioural marker system is comprehensive, coherent, and supported by the wider engineering community (Butler et al., 2020; Hamlet et al., 2023).

Direct observation

While interviews and focus groups provide valuable insights into how experts understand non-technical skills, direct observation allows you to examine how these skills are demonstrated in practice. By observing individuals while they perform authentic tasks, you can identify whether the behaviours described in your emerging framework are visible, meaningful, and relevant in real contexts (Thompson, 2013).

Direct observation is particularly valuable because non-technical skills are often difficult for individuals to articulate. An engineer may recognise effective teamwork, decision making, or situational awareness without being able to explicitly describe the behaviours that demonstrate these skills. Observing performance allows you to capture these behaviours as they occur, reducing reliance on memory and self-reporting. Depending on your context, observations may be conducted during workplace activities, design reviews, laboratory sessions, project meetings, simulations, or other authentic learning activities.

When conducting observations, focus on recording what people do, rather than interpreting their intentions. For example, instead of noting that a student 'demonstrated good communication', capture the specific behaviours that led to this judgement: asked clarifying questions before starting a task, shared relevant information with team members, or adapted their explanation when others showed uncertainty. Similarly, documenting ineffective behaviours is important, as these provide useful contrasts and help define the boundaries of acceptable performance.

In developing a behavioural marker system, direct observation can serve two purposes. First, it can validate whether the behavioural markers identified through literature reviews, interviews, and focus groups are observable in practice. Second, it can reveal new behaviours that were not captured through earlier methods. For example, observing engineers or students working through complex design challenges may highlight subtle behaviours related to collaboration, problem framing, or decision making that experts did not initially articulate (Thompson, 2013). However, direct observation can be challenging to implement. It requires access to appropriate activities, sufficient time for observation, and observers who understand both the technical context and the purpose of the behavioural marker system. In educational settings, you may not be able to observe every student or every activity in depth. A pragmatic approach is therefore to begin with a smaller number of representative observations, using these to test and refine your behavioural markers before scaling up.

A useful approach is to combine observation with the other methods described in this chapter. A literature review can provide the initial structure of your taxonomy; interviews and focus groups can reveal expert knowledge and contextualise behaviours; and direct observation can confirm that these behaviours occur in practice. Using multiple sources of evidence helps ensure that your behavioural marker system is not only theoretically grounded but also reflects the realities of professional practice (Butler et al., 2020).

Validating Behavioural Marker Systems

Once an observational assessment such as a BMS has been developed, it is important to evaluate whether it is reliable, valid, and suitable for its intended purpose. Validation helps to establish confidence that the system provides meaningful and consistent information about non-technical skills, rather than simply reflecting individual assessor interpretation or contextual factors.

Reliability refers to the consistency of assessment results across observers and occasions. In behavioural marker systems, inter-rater reliability is particularly important because it indicates whether different assessors observing the same behaviour apply the rating criteria in a comparable way. A reliable system therefore promotes shared interpretation of performance and limits the extent to which assessment outcomes depend on individual assessor judgement (Fletcher et al., 2003). Test-retest reliability considers whether the system produces consistent results when the same individuals are assessed on different occasions, while internal consistency examines whether the behavioural markers within a particular skill category work together to represent a coherent aspect of performance (Ravindran et al., 2021; Flowerdew et al., 2012).

Validity considers whether the BMS is measuring what it is intended to measure. Content validity is established by ensuring that the framework adequately represents the relevant non-technical skills for the target context, typically through expert review and consultation (Tefera, 2019). Face validity considers whether practitioners recognise the behaviours as meaningful and appropriate, which is important for gaining acceptance and encouraging engagement with the system (Butler et al., 2020). More formal evaluations of construct validity examine whether the BMS captures the underlying skills it aims to assess, while criterion or predictive validity explores whether BMS ratings relate to relevant indicators of performance, such as project outcomes or operational success (Tefera, 2019; Nixon et al., 2015).

For educators implementing a BMS, validation does not need to begin with complex statistical analysis. Practical evaluation methods, such as gathering feedback from assessors and students, reviewing examples of assessed behaviours, and discussing whether ratings align with observed performance, provide valuable evidence for improving the system. Usability and acceptability are particularly important: a BMS must be understandable, manageable to apply, and perceived as useful by those using it (Butler et al., 2020; Thompson, 2013). Finally, consider whether the system is sufficiently sensitive to distinguish between different levels of performance. This is essential if the BMS is intended not only for assessment, but also to support formative feedback and the development of students' non-technical skills (Tefera, 2019).

4 Recommendations for implementation

The successful implementation of observational assessment systems relies on several best practices designed to ensure accuracy, consistency, and meaningful impact. First, rater training is essential for maintaining reliability and fairness. Training should combine theoretical understanding with practical application and calibration exercises so that observers develop a shared interpretation of behaviours and performance standards (Dietz et al., 2014).

The use of video recordings offers additional benefits, enabling repeated viewing for detailed analysis and supporting rater calibration by allowing all observers to assess the same performance. Video also provides concrete examples for feedback and training sessions, helping to ground discussions in observable evidence (Thompson, 2013).

Providing clear feedback and structured debriefing to students is another critical element. Observational systems should support constructive dialogue by linking performance scores to explicit behavioural examples. Debriefing should focus on the implications of behaviours for performance rather than on numerical ratings alone, fostering learning and reflection (Rasmussen et al., 2019; Yule et al., 2008).

A multi-method approach strengthens the validity of findings by combining observational data with complementary sources such as interviews, surveys, or incident analyses. This triangulation provides a more comprehensive picture of performance and its influencing factors (Flin et al., 2017).

Furthermore, iterative development and validation are vital. Your observational framework should evolve over time through systematic testing of reliability, validity, sensitivity, and usability. Long-term validation studies, such as those undertaken to develop BMS in aviation and healthcare, highlight the importance of refining these systems to ensure they accurately differentiate between performance levels (Fletcher et al., 2003, Hamlet et al., 2023).

Finally, organisational support underpins successful implementation. Integrating observational assessment into existing processes requires commitment from leadership as well as buy-in from students and educators (Flin et al., 2017). For an observational framework to become embedded in practice, leadership must recognise non-technical skills as an important component of engineering education and provide the time, resources, and training required for consistent use (which as you have seen from the GENTS-BMS case study, are considerable). This includes supporting educators in interpreting and applying behavioural markers, aligning the assessment approach with existing curriculum and quality processes, and communicating to students that the purpose of the observations is developmental rather than simply evaluative.

References

Ballantyne, I., & Povah, N. (2017). Assessment and development centres. Routledge. https://doi.org/10.4324/9781315262024

Brookfield, S. D. (1995). Becoming a critically reflective teacher. Jossey-Bass.

Butler, P. C., Honey, R. C., & Cohen-Hatton, S. R. (2020). Development of a behavioural marker system for incident command in the UK fire and rescue service: THINCS. Cognition, Technology & Work, 22(1), 1–12. https://doi.org/10.1007/s10111-019-00539-6

Crichton, M. T., Moffat, S., & Crichton, L. (2017). Developing a team behavioural marker framework using observations of simulator-based exercises to improve team effectiveness: A drilling team case study. Simulation & Gaming, 48(3), 299–313. https://doi.org/10.1177/1046878117693266

Dias, R. D., Riley, W., Shann, K., Likosky, D. S., Fitzgerald, D., & Yule, S. (2023). A tool to assess non-technical skills of perfusionists in the cardiac operating room. The Journal of Thoracic and Cardiovascular Surgery, 165(4), 1462–1469. https://doi.org/10.1016/j.jtcvs.2021.06.052

Dietz, A. S., Pronovost, P. J., Benson, K. N., Alejandro Mendez-Tellez, P., Dwyer, C., Wyskiel, R., & Rosen, M. A. (2014). A systematic review of behavioural marker systems in healthcare: What do we know about their attributes, validity and application? BMJ Quality & Safety, 23(12), 1031–1039. https://doi.org/10.1136/bmjqs-2013-002457

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2004). Rating non-technical skills: Developing a behavioural marker system for use in anaesthesia. Cognition, Technology & Work, 6(3), 165–171. https://doi.org/10.1007/s10111-004-0158-y

Flin, R., Yule, S., Paterson-Brown, S., Maran, N., Rowley, D., & Youngson, G. (2006). Experimental evaluation of a behavioural marker system for surgeons' non-technical skills (NOTSS). Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50(10), 969–972. https://doi.org/10.1177/154193120605001036

Flin, R., Martin, L., Goeters, K., Hoermann, J., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (Non-Technical Skills) system for assessing pilots' CRM skills. Human Factors and Aerospace Safety, 3, 95–117. https://www.researchgate.net/publication/224989989

Flin, R., O'Connor, P., & Crichton, M. (2017). Safety at the sharp end: A guide to non-technical skills. CRC Press. https://doi.org/10.1201/9781315607467

Flowerdew, L., Gaunt, A., Spedding, J., Bhargava, A., Brown, R., Vincent, C., & Woloshynowych, M. (2013). A multicentre observational study to evaluate a new tool to assess emergency physicians' non-technical skills. Emergency Medicine Journal, 30(6), 437–443. https://doi.org/10.1136/emermed-2012-201237

Graham, R. (2022). CEEDA case study report: Six case studies of the impact of COVID-19 on global practice in engineering education. MIT. https://www.ceeda.org/wp-content/uploads/CEEDA-report-case-studies.pdf

Hamlet, O. E. D., Irwin, A., Flin, R., & Sedlar, N. (2023). The development of the helicopter non-technical skills (HeliNOTS) behavioural marker systems. Ergonomics 66(12), 2232–2241. https://doi.org/10.1080/00140139.2023.2194591

Hamlet, O. E. D., Irwin, A., & McGregor, M. (2020). Is it all about the mission? Comparing non-technical skills across offshore transport and search and rescue helicopter pilots. The International Journal of Aerospace Psychology, 30(3-4), 215–235. https://doi.org/10.1080/24721840.2020.1803746

Irwin, A., Tone, I. R., & Sedlar, N. (2023). Developing a prototype behavioural marker system for farmer non-technical skills (FLINTS). Journal of Agromedicine, 28(2), 199–207. https://doi.org/10.1080/1059924X.2022.2089420

Jaberi, A., Ciriello, F. & Lucas, C. (2024) Towards Skillful Engineers: A Competence Framework for Engineering Education. UK and Ireland Engineering Education Research Network Conference Proceedings 2023. https://doi.org/10.31273/10.31273/9781911675167/1650

Klampfer, B., Flin, R., Helmreich, R. L., Häusler, R., Sexton, B., Fletcher, G., Field, P. Staender, S., Lauche, K., Dieckmann, P., & Amacher, A. (2001). Enhancing performance in high risk environments: Recommendations for the use of behavioural markers. Gottlieb Daimler & Daimler-Benz Stiftung.

Lucas, C., & Ciriello, F. (2026). GENTS-BMS. [Website]. https://gentsbms.notion.site/

Nixon, J., Leggatt, A., & Campbell, J. (2015). The development and assessment of behavioural markers to support counter-IED training. Applied Ergonomics, 48, 130–137. https://doi.org/10.1016/j.apergo.2014.11.015

O'Connor, P., Hörmann, H. J., Flin, R., Lodge, M., Goeters, K.-M., & The JARTEL Group (2002). Developing a method for evaluating crew resource management skills: A European perspective. The International Journal of Aviation Psychology, 12(3), 263–285. https://doi.org/10.1207/S15327108IJAP1203_5

O'Connor, P., & Max Long, W. (2011). The development of a prototype behavioral marker system for US Navy officers of the deck. Safety Science, 49(10), 1381–1387. https://doi.org/10.1016/j.ssci.2011.05.009

Rasmussen, K., Langdalen, H., Sollid, S. J. M., Abrahamsen, E. B., Sørskår, L. I. K., Bondevik, G. T., & Abrahamsen, H. B. (2019). Training and assessment of non-technical skills in Norwegian helicopter emergency services: A cross-sectional and longitudinal study. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 27(1), Article 1. https://doi.org/10.1186/s13049-018-0583-1

Ravindran, S., Haycock, A., Woolf, K., & Thomas-Gibson, S. (2021). Development and impact of an endoscopic non-technical skills (ENTS) behavioural marker system. BMJ Simulation & Technology Enhanced Learning, 7(1), 17–25. https://doi.org/10.1136/bmjstel-2019-000526

Schrader, P. G., & Lawless, K. A. (2004). The knowledge, attitudes, & behaviors approach: How to evaluate performance and learning in complex environments. Performance Improvement, 43(9), 8–15. https://doi.org/10.1002/pfi.4140430905

Tefera, T. (2019). Behavioural marker systems for evaluation of Norwegian merchant shipping bridge resource management training [Master thesis, Universitetet i Sørøst-Norge]. https://nva.sikt.no/registration/019913c7cec1-02228cd6-70e5-42e1-be91-1fca9658a290

Thompson, D. J. (2013). Behavioural markers of air traffic controller development [Doctoral dissertation, University College London]. https://discovery.ucl.ac.uk/id/eprint/1421081/1/Thompson_1421081_Thesis_nov_2013.pdf

Yule, S., Flin, R., Paterson-Brown, S., Maran, N., & Rowley, D. (2006). Development of a rating system for surgeons' non-technical skills. Medical Education, 40(11), 1098–1104. https://doi.org/10.1111/j.1365-2929.2006.02610.x

Yule, S., Flin, R., Maran, N., Youngson, G., Mitchell, A., Rowley, D., & Paterson-Brown, S. (2008). Debriefing surgeons on non-technical skills (NOTSS). Cognition, Technology & Work, 10(4), 265–274. https://doi.org/10.1007/s10111-007-0085-9

AI statement

During the preparation of this work the authors did not make use of any AI tools.

Contributor statement

Conceptualisation	Francesco Ciriello, Ali Jaber, Claire Lucas
Formal analysis	Francesco Ciriello, Ali Jaber, Claire Lucas
Funding acquisition	Claire Lucas
Investigation	Francesco Ciriello, Ali Jaber, Claire Lucas
Methodology	Ali Jaber
Resources	Ali Jaber
Supervision	Francesco Ciriello, Claire Lucas
Writing - original draft	Francesco Ciriello, Ali Jaber, Claire Lucas
Writing - reviewing and editing	Francesco Ciriello, Ali Jaber, Claire Lucas