This is an open access article under the
Creative Commons Attribution 4.0 International License

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 84, No. 5, 2026
722

GENERATIVE ARTIFICIAL INTELLIGENCE IN CLASSROOMS CALLS FOR OBSERVATION-BASED RUBRICS ALONGSIDE SELF-REPORT SCALES

Servet Demir
Samarkand State University named after Sharof Rashidov, Uzbekistan

Muhammet Usak
University of Kragujevac, Republic of Serbia

Introduction

Generative artificial intelligence (GenAI) is becoming part of teachers' work. In the Teaching and Learning International Survey (TALIS) 2024, which covered 55 education systems, more than one in three teachers across OECD countries (36% on average) reported having used AI in their work, with considerable variation between systems, for example, 14% in France and 75% in Singapore (OECD, 2025). Survey data from the United States point to rapid growth: the share of teachers reporting at least some use of AI was 34% in 2023 and 32% in 2024 and rose to 61% in 2025 (Education Week Research Centre, 2026). These figures come from different populations, years, and definitions of use, and they mainly describe teachers' own use of AI. They nevertheless show that the question of teacher competence in the AI era can no longer be postponed.

The research community has responded quickly. A considerable number of AI-TPACK and Intelligent-TPACK instruments have been developed or adapted in recent years (Celik, 2023; Karataş & Ataç, 2025; Ning et al., 2024), and most of them are self-report scales. Eyal (2025) similarly notes that existing assessments of teachers' AI-related knowledge often rely on self-reporting questionnaires. The argument of this paper is not that these scales are unnecessary. It is that the field now needs, with greater urgency, instruments that capture what teachers actually do with GenAI in classrooms, and that observation-based rubrics are the most important missing piece of this toolkit.

What Self-Report Scales Contribute and Where They Stop

Self-report scales have legitimate strengths. They are economical, they reach large and cross-national samples, and they make structural models of teacher knowledge testable. The Extended GenAI-TPACK framework, on which the present argument builds, itself recommends developing self-report subscales for its new knowledge domains (Demir et al., 2026). However, these considerations are necessary for measurement research; they are not sufficient per se. The TPACK literature shows how the balance between measurement approaches can shift over time. An early review of TPACK measurement categorized 141 instruments into five types that were used fairly evenly, with self-report measures and performance assessments the most frequent (Koehler et al., 2012). A later review of empirical studies found that self-report had become the most common approach and that performance evaluations of teaching activities were rare (Willermark, 2018).

https://doi.org/10.33225/pec/26.84.722 ISSN 1822-7864 (Print) ISSN 2538-7111 (Online)

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 84, No. 5, 2026

The external validity of self-report is a further concern. A recent meta-analysis found that self-reported TPACK showed small-to-moderate correlations with teachers' technology use, with an overall pooled correlation of .32 and dimension-specific correlations of up to .49; part of the use measures were themselves based on teachers' reports (Backfisch et al., 2025). Studies comparing self-reports with performance-based measures point in the same direction. In a study of 141 pre-service mathematics teachers, the correlation between self-reported and test-based technological pedagogical content knowledge was low, r = .24, and was moderated by the teachers' metacognitive accuracy (Fabian et al., 2024). What teachers believe they can do with technology, in other words, explains only part of what they can demonstrate.

There are reasons to expect this gap to matter even more with GenAI, although this remains a hypothesis to be tested. Familiarity with a chatbot interface may easily be mistaken for competence in using it pedagogically. A teacher may rate highly their ability to design AI-supported activities, while the observed lesson shows AI output being presented to students without verification. A questionnaire administered alone is unlikely to reveal such a discrepancy; evidence from practice is needed to do so.

Why Classroom Use of Generative AI Requires Evidence from Practice

Mishra et al. (2023) describe GenAI as protean, opaque and unstable, qualities long attributed to digital technologies in general, and add two qualities that are specific to it: it is generative, producing new content, and social, interacting with its users. When GenAI enters the classroom, interaction is no longer limited to teachers and students but becomes a triangle in which AI can act as an active participant (Demir et al., 2026). Many of the competencies demanded by this triangle are enacted competencies. Detecting and correcting a hallucinated answer in front of the class, deciding which part of a task is delegated to AI and which part remains with the learner, and structuring students' interaction with AI so that it does not turn into cognitive offloading and what Fan et al. (2025) term metacognitive laziness are performances that unfold in time.

The available evidence also indicates that the most demanding forms of classroom AI use are still uncommon. In a RAND survey of K-12 teachers in the United States covering the 2023-2024 school year, one quarter of English language arts, mathematics and science teachers used AI tools for educational planning or teaching; among them, 64% used AI only for planning, and only 36% introduced AI tools to their students at all (Kaufman et al., 2025). A systematic review of 169 studies (2009-2024) on AI teaching methodologies in K-12 and higher education classified 45.56% of implementations as substitutions of human functions and only 4.73% as genuine human-AI partnership (Elsayed et al., 2026). Cukurova et al. (2026), who describe five levels of teacher-AI teaming ranging from transactional to synergistic, found no empirical example of the synergistic level in the teacher-facing studies they reviewed. Classroom use of GenAI with students is therefore at an early stage, and this is precisely the moment at which instruments for observing it should be built, before practices settle and before policy decisions rely on perceptions alone.

Why Observation-Based Rubrics Are Needed

Observation rubrics have a track record in TPACK research. The TPACK-based observation instrument of Hofer et al. (2011) reported an intraclass correlation of .80 for the total score in a small sample of recorded lessons, and the EnTPACK rubric brought lesson planning, implementation, and reflection together in a single instrument (Aumann et al., 2023). Performance-based measures using vignettes and the evaluation of lesson plans have also been developed (Backfisch et al., 2025). An equivalent for the classroom use of GenAI is still missing.

ISSN 1822-7864 (Print) ISSN 2538-7111 (Online) https://doi.org/10.33225/pec/26.84.722

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 84, No. 5, 2026

To the authors' knowledge, the only performance-based AI-TPACK instrument published to date scores written instructional artifacts produced by teacher educators rather than classroom practice, and its author recommends adding classroom observations (Eyal, 2025).

Rubrics for the classroom use of GenAI need to differ from earlier TPACK rubrics in several respects. First, they need observable indicators for competencies that have no counterpart in classical TPACK, such as verifying AI output with students, dividing roles between teacher and AI, managing students' interaction with AI and protecting students' data. Second, they need to judge the appropriateness of a decision rather than the amount of AI used. A teacher who decides not to use AI in a given lesson for sound pedagogical reasons should not be penalized. Third, because the reasons behind a decision cannot be observed directly, classroom evidence needs to be combined with lesson plans and teachers' post-lesson reflections. International frameworks can support the description of performance levels; the AI competency framework for teachers, for instance, organizes fifteen competencies into three progression levels and is intended, among other purposes, to inform the design of assessment (Miao & Cukurova, 2024).

Frameworks that treat AI as an actor in the classroom can complement such rubrics by describing the structure of a lesson. The Hughes Levels × Co-Agency matrix proposed within the Extended GenAI-TPACK framework (Demir et al., 2026) is one such proposal: lessons or lesson segments are coded according to the depth of change in classroom routines and according to whether AI acts directly within the classroom. The matrix has not yet been validated empirically, and its authors describe the binary Co-Agency axis as a conceptual simplification. It is therefore best used as a descriptive profile of AI use, not as a quality score, and alternative coding schemes such as the five teaming levels of Cukurova et al. (2026) deserve to be tested alongside it. Higher cells on such a matrix do not indicate better teaching; appropriateness has to be judged separately.

A Complementary Measurement Design

Self-report scales and classroom rubrics answer different questions and should be designed to support each other. Scales describe perceived competence across large and diverse samples and allow cross-cultural comparison. Classroom rubrics provide evidence of enacted competence in particular lessons. Lesson plans and scenario-based situational judgment tasks occupy an intermediate position: they capture planning and professional judgment, which Demir et al. (2026) suggest may be better assessed with scenario-based tasks than with Likert-type items, but they do not by themselves show what happens in the classroom. Each instrument should therefore be linked explicitly to the inference it is meant to support.

Within such a design, the relation between scale and rubric scores becomes an object of study rather than a validity criterion. Because perceived and enacted competence are expected to differ, only modest correspondence between the two should be anticipated, and the expected pattern of convergent and discriminant relations needs to be specified in advance as part of an argument-based validation. Rubric scores, in turn, bring the obligations of rater-mediated assessment: rater training, anchor exemplars, and analyses that model rater severity, such as many-facet Rasch measurement for rating scales, together with agreement indices for categorical codes such as matrix cells. Generalizability studies that treat lessons as a facet are needed to show how many lessons and raters a dependable judgment requires. Since classroom use of GenAI is still occasional, the sampling of lessons also needs careful planning, as a randomly chosen lesson may contain no AI use at all.

https://doi.org/10.33225/pec/26.84.722 ISSN 1822-7864 (Print) ISSN 2538-7111 (Online)

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 84, No. 5, 2026

Two Cautions

A shift towards observation-based assessment carries two risks that need to be named. The first is technocentrism. A rubric that rewards the amount or sophistication of AI use would penalize the teacher who decides, for sound pedagogical reasons, not to use AI in a given lesson, and would repeat the long-standing tendency to equate more technology with better teaching. The second risk concerns ethics and feasibility. Video recording in classrooms involves minors, data protection, and institutional permissions; prompt histories may also contain personal data; and observation is costly in time. Data protection rules and available resources differ considerably between countries. These constraints are real, but they are an argument for careful design and shared instruments rather than for relying on perceptions alone.

Summing-up

GenAI is changing teachers' work faster than the field is changing the way it measures teachers' competence. Self-report scales inherited from the TPACK tradition will remain useful, and the development of comprehensive scales is part of the same agenda. If research relies on them alone, however, it is likely to describe teachers' perceptions of AI integration in increasing detail while saying comparatively little about the practice of integration. This paper is an attempt to draw attention to the need for observation-based, practice-oriented rubrics that can stand alongside these scales.

A research agenda follows from this argument. Observable indicators for AI-specific competencies need to be defined and subjected to expert review; rubrics need to be calibrated with anchor lessons and rater training; their reliability needs to be examined with rater-mediated psychometric models and generalizability studies; their relations with self-report scales, lesson plans and scenario-based tasks need to be investigated; and their functioning needs to be compared across pre-service and in-service teachers and across countries. It is recommended that journals welcome studies that combine self-report with observational, artifact-based, and scenario-based evidence; that teacher education programs assess pre-service teachers' AI integration in microteaching and practicum settings; and that researchers openly share validated rubrics for classroom use of GenAI.

Declaration of Interest

The authors developed the Extended GenAI-TPACK framework and the Hughes Levels × Co-Agency matrix discussed in this paper and are currently developing a self-report scale and a classroom rubric based on this framework. The authors declare no other competing interest.

Statement on the Use of Generative AI

Claude Opus 5.5 was used to check clarity and proofreading of the manuscript. The authors take full responsibility for the accuracy, integrity, and content of the final manuscript.

References

Aumann, A., Schnebel, S., & Weitzel, H. (2023). The EnTPACK rubric: Development, validation, and reliability of an instrument for measuring pre-service science teachers' enacted TPACK. *Frontiers in Education, 8*, Article 1190152. https://doi.org/10.3389/feduc.2023.1190152

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 84, No. 5, 2026

Backfisch, I., Scherer, R., Schneider, J., & Lachner, A. (2025). How valid, really? A meta-analysis of the validity evidence of technological pedagogical and content knowledge (TPACK) self-report assessments. *Educational Research Review, 49*, Article 100718. https://doi.org/10.1016/j.edurev.2025.100718

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior, 138*, Article 107468. https://doi.org/10.1016/j.chb.2022.107468

Cukurova, M., Suraworachet, W., Zhou, Q., & Bulathwela, S. (2026). Towards synergistic teacher-AI interactions with generative artificial intelligence. In M. C. Mayrath, J. T. Behrens, & D. H. Robinson (Eds.), *Handbook of generative AI in education* (pp. 261-279). Springer. https://doi.org/10.1007/978-3-032-24819-0_8

Demir, S., Uşak, M., & Kodirov, O. S. (2026). Three inseparable facets and five new knowledge domains: An extended GenAI-TPACK proposal. *Pedagogical Research, 11*(3), Article em0268. https://doi.org/10.29333/pr/18714

Education Week Research Center. (2026, January). *More teachers are using AI in their classrooms. Here's why*. Education Week. https://www.edweek.org/technology/more-teachers-are-using-ai-in-their-classrooms-heres-why/2026/01

Elsayed, A. A. A., Malkawi, G., Wardat, Y., & Kamalov, F. (2026). Beyond automation: A systematic review of AI teaching methodologies and a framework for human-AI synergy in higher education. *IEEE Access, 14*, 45948–45963. https://doi.org/10.1109/ACCESS.2026.3675937

Eyal, L. (2025). Developing and validating an AI-TPACK assessment framework: Enhancing teacher educators' professional practice through authentic artifacts. *Education Sciences, 15*(11), Article 1452. https://doi.org/10.3390/educsci15111452

Fabian, A., Fütterer, T., Backfisch, I., Lunowa, E., Paravicini, W., Hübner, N., & Lachner, A. (2024). Unraveling TPACK: Investigating the inherent structure of TPACK from a subject-specific angle using test-based instruments. *Computers & Education, 217*, Article 105040. https://doi.org/10.1016/j.compedu.2024.105040

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Hofer, M., Grandgenett, N., Harris, J., & Swan, K. (2011). Testing a TPACK-based technology integration observation instrument. In M. Koehler & P. Mishra (Eds.), *Proceedings of SITE 2011: Society for Information Technology & Teacher Education International Conference* (pp. 4352–4359). AACE.

Karataş, F., & Ataç, B. A. (2025). When TPACK meets artificial intelligence: Analyzing TPACK and AI-TPACK components through structural equation modelling. *Education and Information Technologies, 30*(7), 8979–9004. https://doi.org/10.1007/s10639-024-13164-2

Kaufman, J. H., Woo, A., Eagan, J., Lee, S., & Kassan, E. B. (2025). *Uneven adoption of artificial intelligence tools among U.S. teachers and principals in the 2023-2024 school year* (Research Report RR-A134-25). RAND Corporation. https://www.rand.org/pubs/research_reports/RRA134-25.html

Koehler, M. J., Shin, T. S., & Mishra, P. (2012). How do we measure TPACK? Let me count the ways. In R. N. Ronau, C. R. Rakes, & M. L. Niess (Eds.), *Educational technology, teacher knowledge, and classroom impact: A research handbook on frameworks and approaches* (pp. 16–31). IGI Global. https://doi.org/10.4018/978-1-60960-750-0.ch002

Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO. https://doi.org/10.54675/ZJTE2084

Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the age of ChatGPT and generative AI. *Journal of Digital Learning in Teacher Education, 39*(4), 235–251. https://doi.org/10.1080/21532974.2023.2247480

Ning, Y., Zhang, C., Xu, B., Zhou, Y., & Wijaya, T. T. (2024). Teachers' AI-TPACK: Exploring the relationship between knowledge elements. *Sustainability, 16*(3), Article 978. https://doi.org/10.3390/su16030978

OECD. (2025). *Results from TALIS 2024: Country notes*. OECD Publishing. https://www.oecd.org/en/publications/2025/10/results-from-talis-2024-country-notes_eafd703e.html

Willermark, S. (2018). Technological pedagogical and content knowledge: A review of empirical studies published from 2011 to 2016. *Journal of Educational Computing Research, 56*(3), 315–343. https://doi.org/10.1177/0735633117713114

Received: *October 06, 2026* Revised: *October 10, 2026* Accepted: *October 25, 2026*

Cite as: Demir, S., & Usak, M. (2026). Generative artificial intelligence in classrooms calls for observation-based rubrics alongside self-report scales. *Problems of Education in the 21st Century*, *84*(5), 722–727. https://doi.org/10.33225/pec/26.84.722

Servet Demir	PhD, Professor, Samarkand State University named after Sharof Rashidov, Samarkand, Uzbekistan. E-mail: servetdemirtr27@gmail.com ORCID: https://orcid.org/0000-0003-1360-2871
Muhammet Usak *(Corresponding author)*	PhD, Associate Professor, Faculty of Education in Jagodina, University of Kragujevac, Republic of Serbia. E-mail: musaktr@gmail.com ORCID: https://orcid.org/0000-0002-6537-9993