caisa Det Nationale Center for AI i Samfundet
The National Center for AI in Society

Fair and Transparent Dialogic Assessment through AI in Higher Education

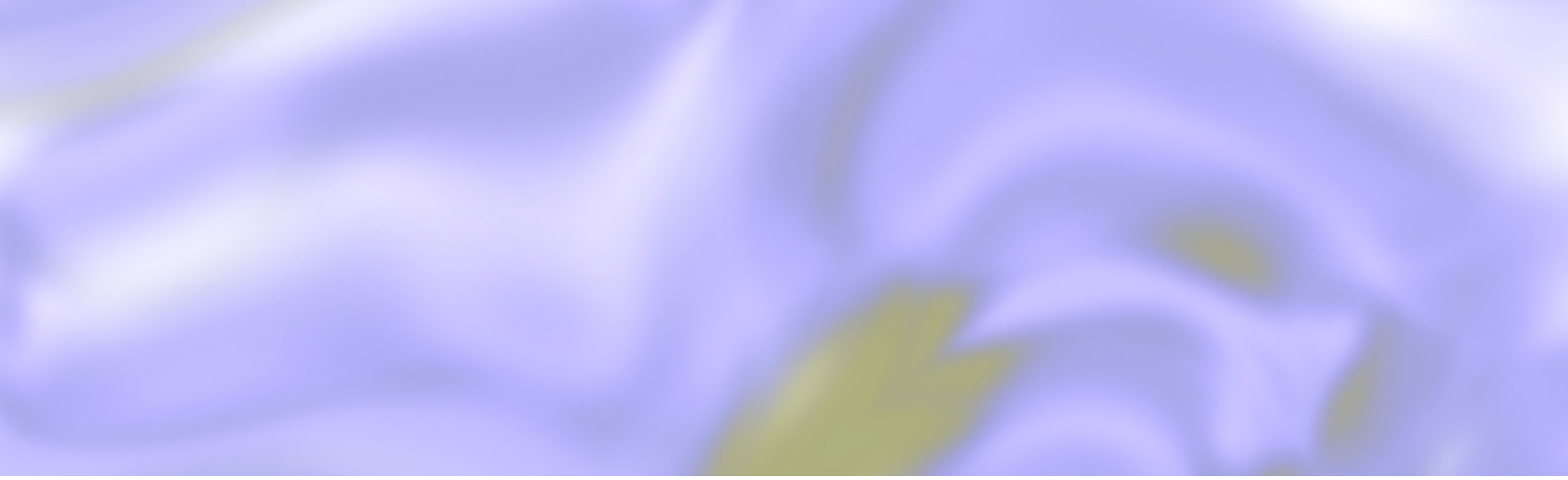
CAISA Research Brief

Authors
Paulsen, Lucas; Davidsen, Jacob; Hontvedt, Magnus; Beal, Charlotte; Steier, Rolf

Editors
Søgaard, Anders; Feldt, Johannes N.

- Publication date: 30 June 2026
- Published by: CAISA, The National Center for AI in Society, Copenhagen and Aalborg, Denmark

- ISSN: 2795-0646
- Document version: Publisher's PDF, registered version
- Citation for published version (APA): Paulsen, L., Davidsen, J., Hontvedt, M., Beal, C. & Steier, R. (2026). Fair and Transparent Dialogic Assessment through AI in Higher Education. *CAISA – Brief.* www.caisa.dk/en/forskning/ai-in-dialogic-assessment-in-higher-education



Fair and Transparent Dialogic Assessment through AI in Higher Education

By Lucas Paulsen, Jacob Davidsen, Magnus Hontvedt, Charlotte Beal and Rolf Steier

Summary

Oral exams and dialogic forms of assessment have gained renewed attention as written assessment is increasingly challenged by generative artificial intelligence (AI). Yet, dialogic assessment remains difficult to document and revisit. Once the exam ends, little remains beyond notes, recollections, and a grade. This brief reports findings from participatory workshops with students and examiners across different faculties at Aalborg University who discussed their experiences with oral exams and reflected on three AI-generated traces of a fictional oral exam: a verbatim transcript, a concept-identification overlay, and a turn-taking visualisation. Students framed traces as a possible basis for accountability, while examiners weighed their value against concerns about access, interpretation and workload. We identify the conditions under which traces can support fair and transparent dialogic assessment, and the decisions that are yet to be explored.

Why dialogic assessment, and why now?

Generative AI has made written assessment increasingly contested, renewing interest in oral formats (Fenton, 2025). In Denmark and Scandinavia, oral exams have long been a central feature of higher education, especially in problem-based learning environments such as Aalborg University (AAU), where group-based project examinations are the primary format (Kolmos & Holgaard, 2009).

However, dialogic assessment carries its own unresolved challenges. When the exam ends, the material basis for revisiting the assessment is often limited – a grade, notes, and recollections of a high-stakes interaction. Research on grading conversations shows that subtle interactional dynamics (who is asked first, how questions are framed) can shape what becomes visible as competence (Solem et al., 2024). Examination is thus not a neutral measurement but a social practice in which examiners and students co-construct what counts as competence (Hontvedt et al., 2024; Wiig, 2023). This can make it difficult for students to understand how the final judgement was reached, particularly in group exams, where individual contributions must be disentangled from collective performance.

Despite decades of pedagogical transformation towards student-centred, active, and dialogue-based learning, the corresponding assessment practices have remained surprisingly static and constrained (Boud & Bearman, 2022). We argue that AI-generated traces of oral exams offer one way to address this opacity – not by replacing human judgement, but by creating shared representations that make the basis for that judgement more fair and transparent (Steier et al., 2026).

AI-generated traces

With partner universities in Norway (University of South-Eastern Norway and Oslo Metropolitan University), we have developed a prototype that uses speech-to-text transcription and large language model analysis to generate structured representations from oral exam recordings (Steier et al., 2026). With students and examiners, we focused on three traces, each making visible a different dimension of a fictional exam.

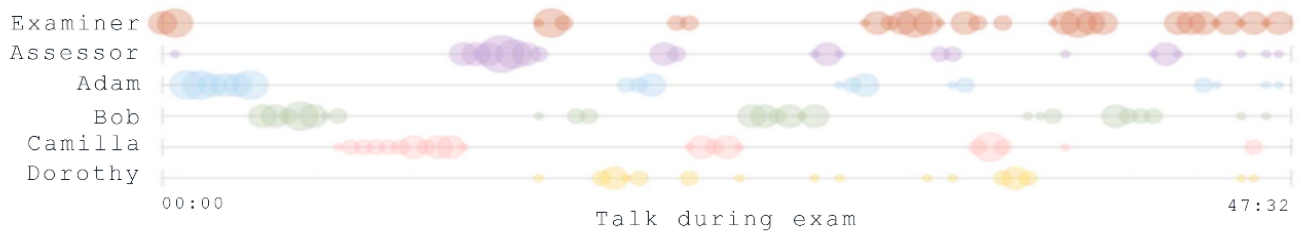
Figure 1. Trace #1 – Verbatim transcript. A time-stamped record of who said what during the exam.

12:34	**Examiner**	Let's discuss organizational culture. Anne, can you explain Schein's model of organizational culture?
12:42	**Anne**	Yes, um... Schein's model has three levels. The most visible level is artifacts — things you can see and touch, like office layout or dress code. Then there's espoused values, which are the stated values and norms. And the deepest level is basic assumptions, which are unconscious beliefs that really guide behavior.
13:08	**Examiner**	Good. Brian, can you elaborate on how these levels interact?
13:14	**Brian**	Sure. The basic assumptions are... they're taken for granted, so people don't question them. They influence espoused values, which then become visible in artifacts. But there can be inconsistencies — like, an organization might say they value innovation in their espoused values, but their basic assumptions might actually favor stability.
13:38	**Examiner**	Exactly. Now, thinking about digital transformation — Camilla, what challenges might arise when an organization tries to change its culture in the context of digitalization?
13:52	**Camilla**	Well, resistance to change is a big challenge. People might resist because the new digital tools conflict with their basic assumptions about how work should be done. Um... Lewin's model talks about unfreezing, changing, and refreezing. So you'd need to unfreeze the old culture first, introduce the new digital practices, and then refreeze them so they become the new normal.
14:26	**Examiner**	Can you give a concrete example of what 'unfreezing' might look like in practice?
14:32	**Camilla**	Uhm... maybe creating a sense of urgency? Like showing employees that the current way isn't sustainable anymore because competitors are digitalizing. Or involving people in the change process so they feel ownership.

Figure 2. Trace #2 – Concept identification. The same transcript but with curriculum-relevant concepts and connections to course literature highlighted through LLM analysis.

12:34	**Examiner**	Let's discuss **organizational culture**. Anne, can you explain **Schein's model** of organizational culture?
12:42	**Anne**	Yes, um... Schein's model has three levels. The most visible level is **artifacts** — things you can see and touch, like office layout or dress code. Then there's **espoused values**, which are the stated values and norms. And the deepest level is basic assumptions, which are unconscious beliefs that really guide behavior.
13:08	**Examiner**	Good. Brian, can you elaborate on how these levels interact?
13:14	**Brian**	Sure. The basic assumptions are... they're taken for granted, so people don't question them. They influence **espoused values**, which then become visible in **artifacts**. But there can be inconsistencies — like, an organization might say they value innovation in their espoused values, but their basic assumptions might actually favor stability.
13:38	**Examiner**	Exactly. Now, thinking about **digital transformation** — Camilla, what challenges might arise when an organization tries to change its culture in the context of digitalization?
13:52	**Camilla**	Well, **resistance to change** is a big challenge. People might resist because the new digital tools conflict with their basic assumptions about how work should be done. Um... **Lewin's model** talks about **unfreezing, changing, and refreezing**. So you'd need to unfreeze the old culture first, introduce the new digital practices, and then refreeze them so they become the new normal.
14:26	**Examiner**	Can you give a concrete example of what 'unfreezing' might look like in practice?
14:32	**Camilla**	Uhm... maybe creating a sense of urgency? Like showing employees that the current way isn't sustainable anymore because competitors are digitalizing. Or involving people in the **change process** so they feel ownership.

Figure 3. Trace #3 – Turn-taking visualisation. A temporal, colour-coded diagram showing the distribution of speaking time across all participants over the duration of the exam.



We want to highlight that traces are designed as decision support, not decision substitutes. Each trace is a representation, not a record of the exam "as it really was." It foregrounds some aspects of the interaction while leaving others invisible. The examiner and assessor remain responsible for interpretation, grading, and feedback.

What students and examiners told us

In spring 2026 we ran three participatory workshops with participants from across AAU's four faculties – two with students and one with examiners. Workshops combined experience-sharing, trace mock-up discussion, and structured ranking of ten implementation dilemmas. Workshops were video-recorded, transcribed, and thematically analysed. The workshops do not show how traces function in actual grading, but they identify the possibilities, concerns and questions that should guide any future pilot.

No single trace is sufficient – the value emerges from the combination

The verbatim transcript drew the most positive response from students. Students described oral examination as a high-stakes situation that is difficult to reconstruct afterwards. A transcript could allow examiners to revisit answers during grading and provide a basis for review in cases of appeal. The transcript's value was framed as protective, but conditional on access. If traces were produced only for examiners, documentation would feel like monitoring, not accountability.

The concept identification trace prompted students to compare the performance of the fictional students in the transcript. It raised questions of whether divergent vocabulary might reflect equivalent or even better understanding, and how "namedropping" concepts differs from knowledge-in-use. Students were more conflicted on the turn-taking visualisation as a stand-alone representation but recognised its value alongside the other two traces.

Importantly, students also identified what traces cannot capture: Whiteboard equations, manipulation of physical artefacts, and gestures like the subtle nodding of an examiner that signals a student is on the right track – aligning with competence as – recognised through embodied interaction, not just talk (Hontvedt et al., 2024; Streeck et al., 2011).

When recall and notes are not enough

The examiner workshop raised several of the same concerns, but from the other side of the exam table. Examiners acknowledged that the "feel" or atmosphere of the room can shape how an exam unfolds, especially during long exam days and large group-based project examinations – the "vibes" that students also perceive as part of assessment. When discussing the traces, examiners described feeling "at the mercy" of the co-examiner's notes, because facilitating discussion and taking detailed notes simultaneously is challenging. Examiners also reflected on how it can be difficult to justify a grade – both to students and to one another – and on how increased documentation may change their approach to questioning during the exam. Traces seemed most relevant where assessment places high demands on recall, note-taking, and individualisation. This includes long group-based project examinations and series of shorter back-to-back exams. Yet examiners also questioned whether traces would be practically usable under real time pressure. With a 25-page group transcript and ten minutes for deciding on a grade, the risk is defaulting to whichever trace is fastest to read rather than engaging with them in combination.

What students and examiners identify as the important questions

In a structured ranking exercise, participants chose from ten paired tensions around the implementation of AI-generated traces. They first voted individually and then negotiated a top three as a group. Two tensions received consistent attention in all workshops.

What traces capture, and what they do not. Participants asked whether traces can support fairness without reducing oral assessment to what can be transcribed, highlighted, or visualised. Students and examiners named similar contextual contingencies – recency bias[1], an examiner who has run several exams already, equations which are written and later erased on a whiteboard. These contextual factors are part of why traces may be useful, but they also define their limits. Institutions should be explicit about which problems traces address and which they cannot. Those limits should shape future trace design, not become an argument against documenting dialogic assessment at all.

Who the traces are for. Students held divergent views on access that reveal the same underlying question. One student framed open student access as basic protection: "there was no protection for us, because we can't really do anything with just the notes from the examiner." Another framed mandatory tracing as disempowering, producing

[1] Where most recent part of the examination weighs heavier than the whole performance in the assessment.

an exam where "everything I'm saying is being surveyed and analysed." Both reflect legitimate student interests.

Examiners were more cautious, noting traces could be read selectively in appeals or detached from context. However, they recognised the underlying asymmetry – that students often have limited access to the material basis needed to understand or formally appeal a grade. The central disagreement is not whether transparency matters, but who gets access to it. Institutions deploying traces must establish how students can access traces, for what purposes, and for how long. Possible models include: (1) examiner-only access during grading, (2) student access in formal appeal cases, or (3) learning-oriented access to selected traces after the appeal window has closed for students and/or examiners.

The third priority differed between the two groups, revealing different concerns on either side of the exam table.

Students emphasised examiner training as a prerequisite. Students argued that without understanding how traces are generated and what they can and cannot represent, AI-generated traces become "a black box on top of a black box". Institutions that deploy traces without examiner training risk embedding the very opacity they set out to reduce.

The examiners emphasised professional trust. Grading conversations rely on notes, criteria, but also on mutual trust in each other's judgement, especially when interpretations differ. Their concern was that traces could shift the character of this conversation, making it more defensive or documentation-driven. The issue is therefore not whether traces should replace professional judgement, but how professional judgement can be made more accountable without undermining the trust and discretion that grading also requires.

Main points

AI-generated traces of oral exams are increasingly technically feasible, but their responsible use is not primarily a technical question. Based on our workshops we foreground the following points for decision-makers in higher education, study boards, and centres for teaching and learning.

- **Oral exams are not automatically fair and transparent.** They can support dialogic, contextualised assessment, but they remain opaque unless fairness is actively supported through assessment design, documentation, examiner training, and student access.
- **Traces should support judgement, not replace it.** Traces are perceived as useful when they let examiners revisit what was said, compare with their notes, and ground grading conversations in shared evidence. They become problematic when they are read as objective measurements or used to delegate judgement. Examiner training is key for ensuring this.
- **No single trace is sufficient.** Transcripts, concept identification, and turn-taking visualisations each make different aspects of the exam visible. Their value lies in combination, alongside examiner judgement and awareness of what traces cannot capture. Yet under time-pressured grading, examiners may default to whichever trace is fastest to read. Supporting their combined value is both a question of trace-design and exam format.
- **Access is a central governance question.** Students often framed access to traces as accountability, while examiners emphasised the need to prevent traces from being interpreted selectively or detached from the wider exam context. Students will be able to request access to traces like they typically can with examiners' notes – but institutions must decide whether traces should be more openly accessible for students, for what purpose, and for how long.
- **Traces act back on assessment practice.** Documenting dialogic assessment is not neutral. Traces can reshape how examiners question and grade, and how students participate in exams. This raises central questions about what documentation means for dialogic assessment as a social practice. While these representations may allow for formative use of summative assessment (using traces for feedback and reflection), they may also make exams more documentation oriented.

What we still need to know before piloting AI-generated traces

In 2026 many of us have participated in meetings or consultations, where AI has been a silent participant. So, we are already facing situations where AI is listening, transcribing and summarising what we say. It is therefore not hard to imagine that AI will be integrated in exams in the future and we need to carefully make decisions about that. Our workshops captured students' and examiners' reactions to mock traces of such AI-documentation in a low-stakes setting. Before piloting, institutions should decide on suitable exam formats, access, storage periods, and examiner training. Pilots should study not only whether traces are technically accurate, but also whether and how they reshape grading conversations, trust, appeals, and workload. The workshops also pointed beyond AI to a broader concern: oral exam practices themselves remain under-researched as social and pedagogical practices. This requires more interactional research on how oral exams unfold, how grading conversations are conducted, and how competence is recognised in dialogue (Hontvedt et al., 2024; Solem et al., 2024; Wiig, 2023). Finally, future

pilots must address questions of digital sovereignty, bias, and interpretation. Even if traces are technically feasible, institutions still need to decide whether AI-generated interpretations of talk can become sufficiently reliable, explainable, and governable for use in assessment.

About the authors

Lucas Paulsen is a Research Assistant, PhD, at the Department of Culture and Communication, Aalborg University. He was a CAISA Fellow in spring 2026.

Jacob Davidsen is an Associate Professor at the Department of Culture and Communication, Aalborg University.

Magnus Hontvedt is a Professor at the Department of Educational Science, University of South-Eastern Norway.

Charlotte Beal is an Associate Professor at the Department of Educational Science, University of South-Eastern Norway.

Rolf Steier is a Professor at the Department of Primary and Secondary Teacher Education, Oslo Metropolitan University.

About CAISA

The National Center for Artificial Intelligence in Society (CAISA) is a national consortium that gathers researchers from the University of Copenhagen, Aalborg University, Aarhus University, the IT University of Copenhagen and the Technical University of Denmark in close collaboration with the Pioneer Centre for Artificial Intelligence (P1).

As Denmark's independent research center for artificial intelligence in society, CAISA centers the citizen. We carry out groundbreaking interdisciplinary research, and we deliver overview of the most recent scientific breakthroughs. Based on new and interdisciplinary research, we advise decision-makers in the public and private sectors on how they may develop and use artificial intelligence practically, such that it contributes to growth, supports democracy and strengthens digital autonomy.

About CAISA's briefs

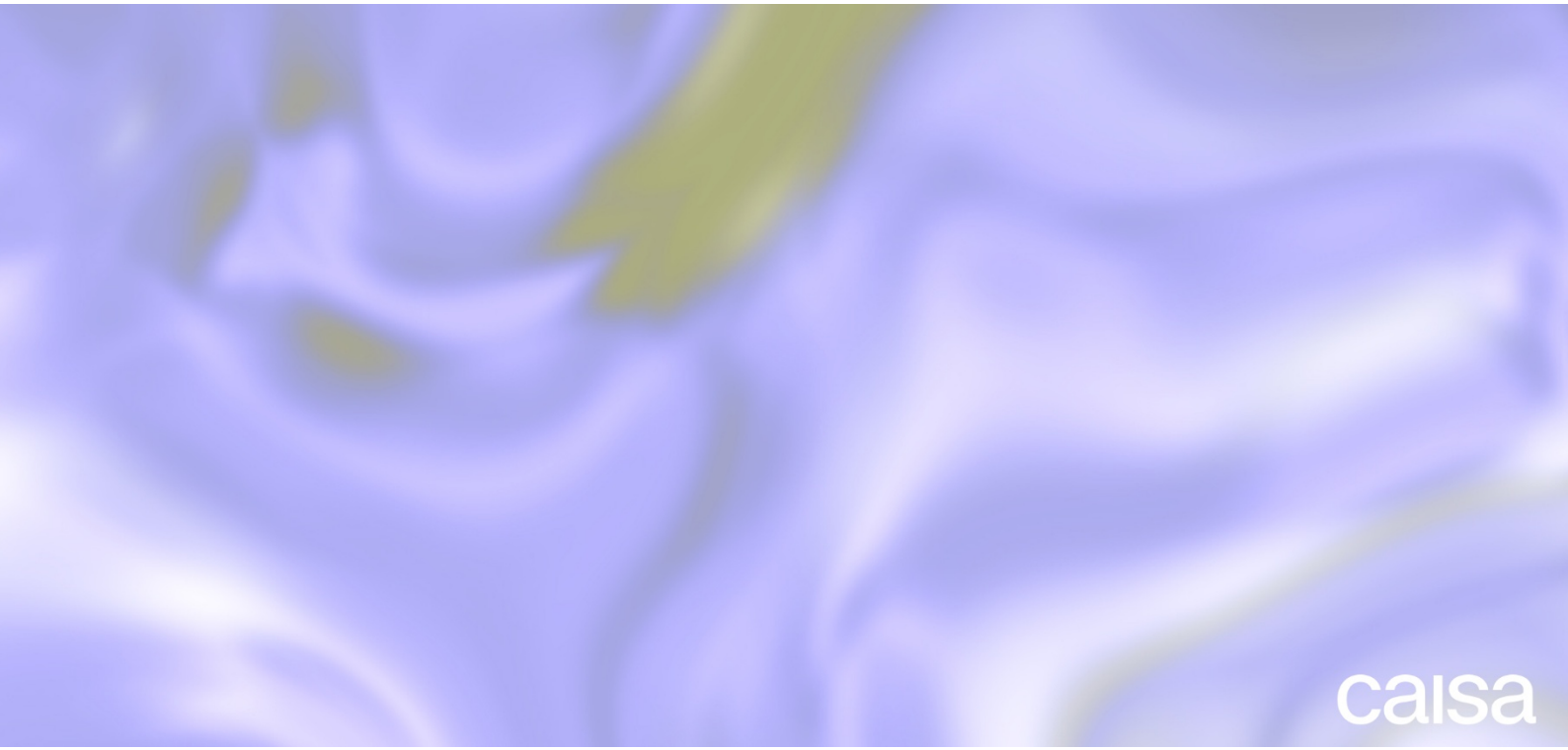
CAISA briefs are part of CAISA's effort to ensure that knowledge and novel insights from within the research community come to empower decision-makers in public and private sectors, and, in turn, society-writ-large when faced with the opportunities and risks posed by rapid technological change. CAISA publishes two types of briefs:

Research briefs present research and evidence-based knowledge about AI and society in an accessible way.

Position briefs express the authors' research based and informed assessment of important challenges related to AI and society.

CAISA's briefs are edited by Anders Søgaard, Professor at the Department of Computer Science at the University of Copenhagen and Chief Scientist at CAISA, and Johannes N. Feldt, Research Assistant at CAISA. All briefs are read by – and receive comments from – at least one external independent researcher before publication.

The authors are responsible for the contents of a CAISA brief.



References

- Boud, D., & Bearman, M. (2022). The assessment challenge of social and collaborative learning in higher education. *Educational Philosophy and Theory, 56*, 1–10. https://doi.org/10.1080/00131857.2022.2114346
- Fenton, A. (2025). Reconsidering the Use of Oral Exams and Assessments: An Old Way to Move Into a New Future. *Educational Researcher, 54*(7), 430–436. https://doi.org/10.3102/0013189X251333638
- Hontvedt, M., Solem, M. S., & Prøitz, T. S. (2024). *Oral Exams as Social Practice: Questions, Joint Remembering, and Co-construction.* https://repository.isls.org//handle/1/11117
- Kolmos, A., & Holgaard, J. E. (2009). Gruppebaseret eller individuel projekteksamen – fordele og ulemper. *Dansk Universitetspædagogisk Tidsskrift, 4*(7), 33–43. https://doi.org/10.7146/dut.v4i7.5595
- Solem, M. S., Landmark, A. M. D., Stokoe, E., & Skovholt, K. (2024). Assessment in practice: Achieving joint decisions in oral examination grading conversations. *Scandinavian Journal of Educational Research, 68*(7), 1522–1539. https://doi.org/10.1080/00313831.2023.2250380
- Steier, R., Hontvedt, M., Davidsen, J., Beal, C., Oddvik, M., & Solem, M. (2026). A Design Space for Dialogic Assessment in CSCL. *Proceedings of the 2026 International Conference on Computer-Supported Collaborative Learning (CSCL)*. ISLS Annual Meeting 2026.
- Streeck, J., Goodwin, C., & Lebaron, C. (2011). Embodied interaction in the material world: An introduction. In J. Streeck, C. Goodwin, & C. Lebaron (Eds.), *Embodied Interaction: Language and Body in the Material World* (pp. 1–26). Cambridge University Press.
- Wiig, A. C. (2023). *Tracing Policy in Practice. Exploring the Interactional Exercise of Oral Assessment* (pp. 265–285). https://doi.org/10.1007/978-3-031-36970-4_14

CAISA's website

CAISA's LinkedIn