

ADAPTING THE DYNAMIC MODEL OF EDUCATIONAL EFFECTIVENESS FOR TEXT-BASED ANALYSIS – CODEBOOK DEVELOPMENT AND CONTENT VALIDATION FRAMEWORK

J. Vaitekaitis^{1,3}, L. Kyriakides², M. Lukoševičius³

¹*Vilnius University (LITHUANIA)*

²*University of Cyprus (CYPRUS)*

³*Kaunas University of Technology (LITHUANIA)*

Abstract

As lesson transcripts become increasingly accessible, there is a growing need for tools that can capture teaching quality from text alone, laying the foundation for scalable, AI-driven analysis. This study adapts the classroom-level components of the Dynamic Model of Educational Effectiveness (DMEE) into a low-inference codebook specifically optimized for textual data. The goal is to preserve the theoretical integrity of the original in-class observation instruments while ensuring that each indicator is operationalized through explicit verbal evidence, creating a robust framework suitable for both human annotation and future machine learning applications.

We started by analyzing all existing elements of low- and high-inference observation instruments, adding new sub-items to capture characteristics specific to classroom transcripts. Each element was reviewed to determine its dependence on nonverbal cues, precise timing, or holistic assessment. The elements were then classified according to a systematic discernibility scheme: textually discernible, conditionally discernible, impossible to discern from the text, or omitted due to redundancy. Where possible, complex indicators were reformulated into smaller verbal events so that the main construct could still be presented in a concise manner.

This article describes in detail the methodological framework used to develop the final instrument. To establish content validity, we employed a structured expert panel review in which experts in the DMEE independently assessed the relevance and clarity of each indicator. The analysis showed high content validity, and only minor improvements were needed based on expert feedback. In the validated codebook, more than a hundred indicators are grouped into conceptual levels that distinguish simple turn-level codes from context-dependent or derived metrics.

The developed framework offers researchers and practitioners a DMEE theory-grounded, content-validated approach for analysing lesson transcripts at scale. It clearly shows which aspects of effective teaching can be captured from text alone and which remain poorly represented, thus defining a realistic scope for text-based classroom observation and laying the groundwork for subsequent reliability testing and semi-automated analysis.

Keywords: Dynamic Model of Educational Effectiveness, classroom observation, transcript analysis, text-based analysis, codebook development, content validity, teacher evaluation.

1 INTRODUCTION

Classroom observation remains a central method for teacher evaluation and instructional improvement, yet it faces persistent scalability constraints, including high resource demands and reliability challenges when observations are conducted by school personnel or evaluators with limited calibration [1], [2]. At the same time, the availability of classroom dialogue transcripts has increased through large educational corpora and improved transcription pipelines, enabling repeated asynchronous review and opening new opportunities for scalable analysis [3].

However, established observation frameworks-particularly those designed for live or video observation - often assume access to multimodal evidence such as nonverbal cues, physical context, and holistic impressions. This creates a methodological barrier when researchers attempt to measure instructional quality from text alone. The Dynamic Model of Educational Effectiveness (DMEE) provides one of the most comprehensive and theoretically grounded models linking teaching practices to student outcomes, operationalized via detailed observation instruments [4], [5]. Yet, transferring DMEE instruments to

transcripts requires careful adaptation so that indicators remain faithful to their constructs while being measurable through explicit verbal evidence only.

This paper addresses this gap by presenting a methodological framework for adapting classroom-level DMEE components to text-based analysis and by reporting content validity evidence produced through expert review. The purpose is not to claim that text can replace multimodal observation; rather, it is to define a transparent and theory-aligned measurement scope that is feasible in transcript-only data and suitable for future AI-supported analytics [6], [7].

2 METHODOLOGY

2.1 Design principles for text-only adaptation

The adaptation was guided by three principles.

Low-inference constraint. Indicators must be triggered by explicit linguistic evidence (verbal acts, explicit statements, or clearly defined dialogue sequences). Items that require interpreting internal states (e.g., “students understood”), nonverbal behaviour (e.g., “students on-task”), or global classroom climate without textual evidence are treated as incompatible with text-only coding.

Discernability-based filtering. Each candidate indicator is evaluated for whether it can be identified in transcripts with acceptable clarity. This is operationalized as a discernability scheme that separates directly observable verbal events from items that are only measurable under additional conditions (e.g., reliable timestamps) and from items that are fundamentally unmeasurable from text.

Construct-preserving reformulation. When an original DMEE item represents an important construct but is too high-inference as written, the item is reformulated into a smaller, verifiable textual proxy (for example, by shifting from “students understood X” to “a student accurately articulates X after the teacher’s explanation”).

2.2 Item pool and adaptation workflow

The starting pool was derived from DMEE low- and high-inference observation instruments and extended with transcript-specific sub-items to capture phenomena visible in classroom talk but not well represented in the original schedules [4], [5]. Each candidate indicator was then processed through the following workflow:

- 1 **Dependency check.** Does the indicator depend on (a) nonverbal cues, (b) precise timing unavailable in transcripts, or (c) holistic judgements not evidenced by text?
- 2 **Classification.** Each indicator is classified as:
 - **Discernable:** can be coded from explicit verbal evidence;
 - **Conditionally discernable:** can be coded if specific conditions are met (e.g., timestamps, explicit verbal marking);
 - **Impossible:** not measurable from text due to modality/inference dependence;
 - **Omitted:** measurable but excluded due to operational redundancy (e.g., duplicate textual evidence).
- 3 **Reformulation (where feasible).** High-inference items are transformed into low-inference proxies anchored in turn-level verbal evidence or defined multi-turn sequences.

2.3 Operational structuring for practical use

Transcript-based instruments often become large, creating cognitive burden for coders and increasing inconsistency. To mitigate this risk, the codebook is organized into conceptual operational levels:

- **Turn-level indicators** (direct manual codes): discrete verbal acts (e.g., explicit orientation statements, question types, feedback moves);
- **Context-dependent indicators:** codes requiring immediate dialogue context;
- **Derived or sequence indicators:** items that are more reliably computed via scripts (multi-turn patterns and timestamp-based metrics), particularly important when transcript metadata supports temporal analysis [8].

This structuring is a practical operational decision intended to support later reliability testing rather than a claim that reliability has already been demonstrated.

2.4 Content validity via structured expert review

To evaluate whether the adapted indicators preserve the theoretical intent of DMEE and are sufficiently clear for consistent application, a structured expert panel review was conducted. A panel of five subject matter experts was recruited to review the instrument. To ensure rigorous evaluation, strict inclusion criteria were applied: panelists were required to possess (a) advanced academic credentials (PhD or doctoral candidacy) in educational effectiveness research, and (b) specific, demonstrated experience in using or developing the original DMEE observation instruments. This ensured the panel could accurately judge whether the text-based proxies maintained construct validity. Experts rated each indicator on standard content validity dimensions, focusing on:

- **Relevance:** alignment of the text-based indicator with the intended DMEE construct;
- **Clarity:** whether the operational definition is sufficiently clear to support consistent coding.

Scale-level indices indicated strong agreement, and feedback was used to implement targeted refinements (e.g., clarifying boundaries between overlapping indicators and strengthening inclusion/ exclusion rules).

3 RESULTS

3.1 Content validity outcomes

The expert review produced **high content validity** across the adapted instrument, with only **minor refinements** required based on panel feedback. Improvements were primarily directed toward: (a) sharpening boundaries between semantically adjacent categories (e.g., procedural directions vs. task/exercising instructions), and (b) specifying more explicit evidence conditions for ambiguous cases. These findings support the conclusion that the instrument is both theoretically grounded in DMEE and sufficiently well-defined to proceed to reliability testing.

3.2 Differential feasibility across DMEE factors

A central outcome of the adaptation is **uneven feasibility** across DMEE's eight classroom-level factors when restricted to text-only evidence. Factors enacted primarily through discourse (e.g., questioning, feedback, explicit structuring and orientation talk) are strongly represented in transcripts. In contrast, constructs depending on nonverbal climate, silent engagement, or broader assessment cycles remain only partially measurable and are captured, at best, via narrow verbal proxies. This aligns with prior work showing both the richness of classroom dialogue and the need for careful coding frameworks to interpret it across contexts [9], [10].

Table 1 summarizes this feasibility pattern qualitatively, emphasizing that feasibility does not equal completeness of construct representation.

Table 1. Qualitative feasibility of DMEE factors in text-only transcripts.

<i>DMEE factor</i>	<i>Text-only feasibility</i>	<i>Main reason</i>
Orientation	High	Often explicitly verbalized (goals, relevance, framing)
Structuring	Moderate	Many verbal markers exist, but coherence/flow is partly nonverbal/ holistic
Questioning techniques	High	Questions, probing, and feedback are inherently textual
Teaching–modelling	Moderate–High	Modelling talk is observable; success/effect requires cautious proxies
Application/exercising	Moderate	Task setup is verbal; student engagement during work is mostly nonverbal
Management of time	Conditional	Strongly improved by timestamps; weak without temporal metadata
Classroom learning environment	Low–Moderate	Verbal traces exist, but overall climate and silent engagement are not visible
Assessment practices	Low	Many assessment components occur outside text; verbal elements overlap with questioning/feedback

3.3 What the text-based instrument can and cannot claim

The adapted codebook provides a valid approach for analysing **discursive enactments** of teaching—what teachers and students say and how interactions unfold in dialogue. It does not claim to measure the full richness of multimodal classroom life. In practical terms:

- The instrument is well suited for identifying **instructional talk patterns** (e.g., question focus, feedback moves, explicit structuring).
- It is less suited for measuring constructs where the core evidence is **silent, nonverbal, or contextual** (e.g., general engagement, classroom emotional tone, many assessment activities).
- The availability of timestamps can substantially strengthen indicators related to pacing, pauses, and interaction proportions, converting certain otherwise weak domains into quantifiable metrics.

3.4 Implications for scalable and AI-assisted transcript analysis

The resulting framework supports scalable research uses such as corpus-level description of instructional talk, comparison of instructional patterns across settings, and the creation of curated benchmark data for automated analysis. Recent work evaluating language models for classroom dialogue and instruction-quality measurement underscores the importance of well-defined human-validated instruments as benchmarks, rather than relying on model judgments alone [6], [7].

4 CONCLUSIONS

This paper presented a methodological framework for adapting the classroom-level components of the Dynamic Model of Educational Effectiveness (DMEE) for use with text-only lesson transcripts. The approach addresses a central constraint of transcript-based observation: many established DMEE indicators presuppose multimodal evidence (e.g., nonverbal engagement, classroom climate, or holistic judgements), which cannot be recovered reliably from text alone. To preserve theoretical alignment while maintaining low-inference measurement, candidate indicators were systematically screened for modality dependence, classified using a discernability scheme, and—where feasible—reformulated into explicit textual proxies anchored in observable verbal evidence.

A structured expert panel review provided evidence of strong content validity for the adapted instrument, with feedback primarily leading to refinements that strengthened boundary rules between adjacent categories and clarified evidence thresholds for ambiguous cases. At the same time, the feasibility analysis underscored that DMEE transferability to text is uneven across factors: discourse-enacted dimensions (e.g., questioning, feedback, orientation and structuring moves) are comparatively well represented, whereas factors relying heavily on nonverbal cues or broader assessment cycles are only partially captured through narrow verbal proxies. Importantly, the availability of timestamps substantially increases the analytic potential of transcript data by enabling derived metrics related to pacing, pauses, and interaction proportions.

The contribution of this work is therefore twofold: (1) it provides a theory-grounded, content-validated pathway for analysing DMEE-relevant teaching practices in transcripts at scale; and (2) it clearly delineates the scope and limitations of what text-only classroom observation can claim. Future work will focus on formal reliability testing with independent coders and on the development of curated benchmark datasets to support semi-automated analysis, including evaluation of large language models against human-coded reference standards.

ACKNOWLEDGEMENTS

This research was funded by the Research Council of Lithuania (application No. P-PD-23-133).

REFERENCES

- [1] A. D. Ho and T. J. Kane, “The reliability of classroom observations by school personnel (MET Project: Research Paper),” Bill & Melinda Gates Foundation, 2013.
- [2] M. A. Kraft and A. F. Gilmour, “Can principals promote teacher development as evaluators? A case study of principals’ views and experiences,” *Educational Administration Quarterly*, vol. 52, no. 5, pp. 711–753, 2016.

- [3] D. Demszky and H. Hill, "The NCTE transcripts: A dataset of elementary math classroom transcripts," arXiv preprint, arXiv:2211.11772, 2022.
- [4] B. P. M. Creemers and L. Kyriakides, *The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice, and Theory in Contemporary Schools*. Routledge, 2008.
- [5] L. Kyriakides, B. P. M. Creemers, and A. Panayiotou, *Quality and Equity in Education: Revisiting Theory and Research on Educational Effectiveness and Improvement*. Springer, 2020.
- [6] Y. Long, H. Luo, and Y. Zhang, "Evaluating large language models in analysing classroom dialogue," *npj Science of Learning*, vol. 9, no. 1, Art. 60, 2024.
- [7] P. Xu, J. Liu, N. Jones, J. Cohen, and W. Ai, "The Promises and Pitfalls of Using Language Models to Measure Instruction Quality in Education," *EdWorkingPaper No. 24-948*, Annenberg Institute for School Reform at Brown University, 2024.
- [8] D. Wang and G. Chen, "Are perfect transcripts necessary when we analyze classroom dialogue using AIoT?" *Internet of Things*, vol. 25, Art. 101105, 2024.
- [9] S. Hennessey et al., "Developing a coding scheme for analysing classroom dialogue across educational contexts," *Learning, Culture and Social Interaction*, vol. 9, pp. 16–44, 2016.
- [10] Y. Tao and G. Chen, "Coding schemes and analytic indicators for dialogic teaching: A systematic review of the literature," *Learning, Culture and Social Interaction*, vol. 39, Art. 100702, 2023.