



Designing Classroom Observation Rubrics That Capture Dialogic Teaching in Large Classes

Mateo Natarajan^{1*}, Arjun Henriques², Seojun Barbosa³

¹Department of Social Sciences, Pannonian Center for Social Sciences, Mexico

²Department of Social Sciences, Andes Research Institute of Social Sciences, Croatia

³Department of Social Sciences, Mekong Institute for Advanced Social Sciences, Jamaica

Article Info

Volume 1, Issue 1, 2026

[doi: 10.62715/ijetl.2026.v1.i1.pp19-24](https://doi.org/10.62715/ijetl.2026.v1.i1.pp19-24)

Received: 24 November 2025

Accepted: 06 February 2026

Published: 14 April 2026

Abstract

Observation instruments for dialogic teaching were mostly developed in classrooms of twenty to thirty students, where an observer can hear nearly everything that is said. In classes of sixty or more, which are normal in many school systems and in university lecture courses, the same instruments misfire. Whole-class exchanges involve a small share of the students, much of the talk that matters happens in pairs and groups out of the observer's hearing, and teachers who work skilfully within these limits can be scored as monologic. This article proposes a procedure for designing an observation rubric suited to such settings. It moves through six stages: defining the construct, choosing the unit of observation, writing indicators that an observer can actually perceive in a large room, building level descriptors, planning how lessons and groups will be sampled, and assembling evidence of reliability and validity. For each stage the article explains the design choice, the reasoning from the research on classroom dialogue and on observation measurement, and the adaptation that class size requires. The procedure is a proposal. No rubric built with it has been trialled for this article, and the final section sets out what a validation study would have to show before scores were used for any purpose beyond professional discussion.

Keywords: *Dialogic teaching, Classroom observation, Rubric design, Large classes, Classroom talk, Observation reliability*

© 2026 Mateo Natarajan et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

Classroom talk has been studied systematically for half a century, and one finding has been stable throughout. The default pattern of lessons is a three-part exchange in which the teacher asks a question with a known answer, a student replies briefly, and the teacher evaluates the reply (Sinclair and Coulthard, 1975; Mehan, 1979). Dialogic teaching is the name given to practice that departs from this pattern: questions that are open, answers that are extended and built upon, and students who respond to one another as well as to the teacher (Alexander, 2008). Interest in it has grown as evidence has accumulated that some features of dialogue are associated with student attainment (Howe et al., 2019).

* Corresponding author: Mateo Natarajan, Department of Social Sciences, Pannonian Center for Social Sciences, Mexico. E-mail: mateo.natarajan@pannonian.edu

© 2026 Mateo Natarajan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To develop such teaching, or to study it, one must be able to see it reliably, and that means an observation instrument. Detailed coding schemes exist, of which the Scheme for Educational Dialogue Analysis is the most thoroughly documented (Hennessy *et al.*, 2016). They work from transcripts of recorded lessons and assume that most talk is audible. General observation systems such as the Classroom Assessment Scoring System rate instructional dialogue as one dimension among many (Pianta *et al.*, 2008). Neither kind was designed for a room of seventy students on fixed benches, where a single observer hears the teacher and the front rows and little else.

This article sets out how a rubric for dialogic teaching can be designed with large classes in view from the start. It is addressed to teacher educators, school inspectors and researchers who need an instrument that one trained observer can use live. The procedure is proposed and argued for. It has not been used to build and test a rubric here.

2. Background

The design draws on two literatures that rarely meet: research on what dialogic teaching is, and measurement research on what makes an observation score trustworthy.

2.1. What counts as dialogic

Definitions overlap without coinciding. Alexander (2008) characterises dialogic teaching by principles: it is collective, reciprocal, supportive, cumulative and purposeful. Nystrand *et al.* (1997) operationalised dialogic instruction through authentic questions, to which the teacher does not hold a prescribed answer, and uptake, in which a question incorporates what a student has just said. Mercer and Littleton (2007) describe exploratory talk among students, in which reasons are given and challenged. The accountable talk framework specifies accountability to the learning community, to standards of reasoning and to knowledge (Michaels *et al.*, 2008). A review of four decades of research concluded that the field has described patterns of dialogue far more than it has measured their consequences (Howe and Abedin, 2013). A rubric designer must therefore choose among accounts and say which has been chosen.

2.2. Why class size changes the problem

In a large class the arithmetic of turn-taking sets a ceiling on whole-class dialogue. If one person speaks at a time, most students will not speak in a given lesson however well the teacher questions. Teachers respond with choral answers, with brief pair talk, and with group tasks followed by reports from a few groups. Studies of classroom interaction in sub-Saharan Africa, where large classes are common, describe lessons dominated by recitation and chorus and argue that realistic improvement has to start from those conditions (Hardman *et al.*, 2008). An instrument that rewards only extended whole-class exchanges will place nearly every large-class lesson at the bottom of its scale and will fail to distinguish a teacher who has students reasoning in pairs from one who lectures.

2.3. What measurement research requires

An observation score varies with the teacher, and also with the lesson, the observer and the segment observed. Hill *et al.* (2012) showed that agreement between raters is an incomplete test of an observation system and that a generalisability study is needed to learn how many lessons and raters a dependable score requires. Validity is treated in current thinking as an argument: the intended use of scores is stated, and evidence is assembled for each inference on which that use depends (Kane, 2006). Bell *et al.* (2012) applied this approach to observation protocols. Reviews of scoring rubrics add that analytic rubrics with clear descriptors, exemplars and rater training are more reliable than holistic judgements (Jonsson and Svingby, 2007).

3. The Proposed Design Procedure

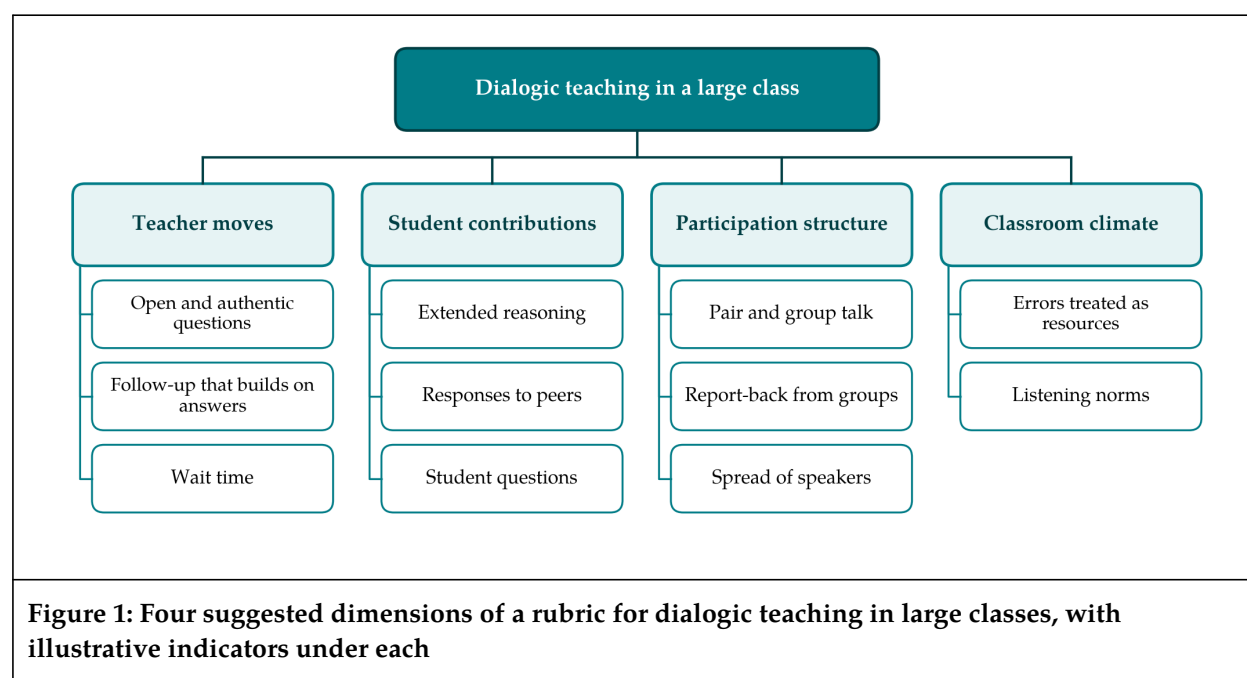
The procedure has six stages. Table 1 summarises each with the principal design decision and the adaptation that a large class calls for. The stages are sequential, though a pilot will normally send the designer back to revise indicators and descriptors.

Table 1: Stages of the proposed rubric design procedure and the adaptation each requires for large classes

Stage	Design decision	Adaptation for large classes
1. Define the construct	Select and state a definition of dialogic teaching	Include pair and group talk within the definition
2. Unit of observation	Decide on time segments or lesson episodes	Separate whole-class segments from group-work segments
3. Write indicators	List behaviours an observer can see or hear	Drop indicators that need every voice to be audible
4. Level descriptors	Describe three or four levels per dimension	Anchor levels in quality and spread, not in number of speakers
5. Sampling plan	Fix lessons, segments and observer position	Sample groups by rule; move during group work
6. Reliability and validity	Plan rater training and the evidence to collect	Check that scores do not simply mirror class size

3.1. Stage 1: Define the construct

The designer writes a definition of one paragraph and derives from it a small number of dimensions. Four are suggested here, shown with illustrative indicators in Figure 1: teacher moves, student contributions, participation structure and classroom climate. The third dimension is the one that conventional instruments lack. It asks whether the lesson is organised so that many students have the chance to talk about the subject matter, and whether what is said in pairs and groups is brought back into the public discussion. Including it makes structured peer talk part of the construct, consistent with the place of exploratory talk in the research ([Mercer and Littleton, 2007](#)).



3.2. Stage 2: Choose the unit of observation

Classical interaction analysis codes classroom events at intervals of a few seconds, which yields fine-grained data and exhausts a live observer. A rubric is better served by longer segments, each rated once. The proposal is to divide the lesson into segments of roughly ten minutes and to label each by its organisation: whole-class teaching, pair or group work, or individual seatwork. Dimensions are rated

only where they apply. This prevents a lesson with substantial group work from being penalised on indicators that presuppose whole-class exchange, and it produces a simple profile of how lesson time was used.

3.3. Stage 3: Write observable indicators

Each indicator must pass a practical test: an observer standing at the side or back of a crowded room, without a transcript, has to be able to tell whether it occurred. Indicators concerning the teacher usually pass, since the teacher is audible. Whether a question is open, whether a follow-up builds on the student's answer, and whether the teacher restates a contribution so that the class can respond to it, the move described as revoicing by O'Connor and Michaels (1993), can all be judged live. Indicators concerning students need rewording. "Students give extended reasons" becomes observable if it is tied to the contributions made publicly and, during group work, to the groups the observer is sampling. Indicators that depend on hearing every student should be removed, and the omission recorded as a limit of the instrument.

3.4. Stage 4: Build level descriptors

For each dimension the designer writes three or four levels, each a short description of what the observer would see. Two cautions apply. Levels should describe quality and spread and should avoid raw counts, because the number of students who speak is bounded by class size and lesson length. A descriptor such as "contributions come from different parts of the room and include students who did not volunteer" captures spread without a count. And the top level should describe practice that a capable teacher can achieve in a large class. If no lesson in the intended setting could reach it, the scale has a level nobody will use, and the remaining levels will discriminate poorly.

3.5. Stage 5: Plan the sampling of lessons and groups

Three sampling decisions are written into the observer's manual. The first is how many lessons are observed for each teacher and across which topics, since a single lesson is a weak basis for any statement about a teacher's practice. The second is where the observer is placed during whole-class teaching, ideally where both the teacher and a cross-section of students can be heard. The third, particular to large classes, is how groups are chosen during group work. Left to choice, observers drift towards the nearest or liveliest group. A rule fixed in advance, for example the group nearest each of three predetermined points in the room for a set number of minutes each, gives a defensible sample. An audio recorder on a sampled desk can supplement live notes where consent and resources allow.

3.6. Stage 6: Assemble reliability and validity evidence

Observer training uses video or live lessons scored independently and then discussed until descriptors are applied consistently. Agreement is reported with a chance-corrected statistic such as kappa (Cohen, 1960), and a small generalisability study estimates how much score variance is due to teachers, lessons, raters and segments (Hill et al., 2012). The validity argument follows Kane (2006). Experts in classroom dialogue review whether the indicators represent the construct; rubric scores for a subset of lessons are compared with detailed coding of the same lessons from recordings; and the relation between scores and class size is examined, because a rubric whose scores fall steadily as classes grow is measuring conditions more than teaching.

4. Fitting the Procedure to Its Purpose

How far each stage is taken depends on what the rubric is for. An instrument for peer observation and professional conversation can tolerate lower reliability, since its value lies in the discussion it prompts, and it benefits from being short. An instrument for research needs the full generalisability and validity work. An instrument that will feed into the appraisal of teachers carries the heaviest burden of evidence, and the designer should be wary of the use: dialogic practice is hard to produce on demand, and high-stakes observation invites staged lessons. Alexander's later work treats repertoire, the range of kinds of talk a teacher can draw on, as the aim (Alexander, 2020), and a rubric used formatively can describe

that repertoire without ranking teachers.

5. Limitations and What Remains to Be Shown

No rubric has been constructed and tested with this procedure for the present article. Its claims are design arguments, and several could fail in practice. Live observers may not be able to judge uptake and revoicing reliably in noisy rooms. Sampled groups may not represent the class. The participation dimension may reward the mere presence of group work whatever its quality. A validation study should therefore report rater agreement by dimension, the number of lessons needed for a dependable score, the correspondence between live ratings and transcript-based coding, and whether scores respond to professional development aimed at classroom talk.

A broader limitation concerns language. In many large-class settings lessons move between an official language of instruction and the languages students speak at home, and dialogue may be richest in the latter. An observer who does not understand both will miss it, and the procedure does not solve that problem beyond recommending that observers share the classroom's languages.

6. Conclusion

Rubrics for dialogic teaching have assumed small classes, and when applied to large ones they tend to record the constraint and overlook the teaching. The six-stage procedure proposed here builds class size into the instrument: the construct includes structured peer talk, segments are distinguished by organisation, indicators are limited to what an observer can perceive, descriptors refer to spread and quality, and groups are sampled by rule. These choices are reasoned from the literature on classroom dialogue and on observation measurement. Their worth will be known only when rubrics designed in this way have been put through the validation the final stage describes.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Alexander, R. (2020). *A dialogic teaching companion*. Routledge, London.
- Alexander, R.J. (2008). *Towards dialogic teaching: Rethinking classroom talk (4th ed.)*. Dialogos, York.
- Bell, C.A., Gitomer, D.H., McCaffrey, D.F., Hamre, B.K., Pianta, R.C. and Qi, Y. (2012). *An argument approach to observation protocol validity*. *Educational Assessment*, 17(2-3): 62-87.
- Cohen, J. (1960). *A coefficient of agreement for nominal scales*. *Educational and Psychological Measurement*, 20(1): 37-46. <https://doi.org/10.1177/001316446002000104>
- Hardman, F., Abd-Kadir, J. and Smith, F. (2008). *Pedagogical renewal: Improving the quality of classroom interaction in Nigerian primary schools*. *International Journal of Educational Development*, 28(1): 55-69.
- Hennessy, S., Rojas-Drummond, S., Higham, R., Márquez, A.M., Maine, F., Ríos, R.M., García-Carrión, R., Torreblanca, O. and Barrera, M.J. (2016). *Developing a coding scheme for analysing classroom dialogue across educational contexts*. *Learning, Culture and Social Interaction*, 9: 16-44. <https://doi.org/10.1016/j.lcsi.2015.12.001>
- Hill, H.C., Charalambous, C.Y. and Kraft, M.A. (2012). *When rater reliability is not enough: Teacher observation systems and a case for the generalizability study*. *Educational Researcher*, 41(2): 56-64. <https://doi.org/10.3102/0013189X12437203>
- Howe, C. and Abedin, M. (2013). *Classroom dialogue: A systematic review across four decades of*

- research. *Cambridge Journal of Education*, 43(3): 325-356. <https://doi.org/10.1080/0305764X.2013.786024>
- Howe, C., Hennessy, S., Mercer, N., Vrikki, M. and Wheatley, L. (2019). *Teacher-student dialogue during classroom teaching: Does it really impact on student outcomes?* *Journal of the Learning Sciences*, 28(4-5): 462-512. <https://doi.org/10.1080/10508406.2019.1573730>
- Jonsson, A. and Svingby, G. (2007). *The use of scoring rubrics: Reliability, validity and educational consequences.* *Educational Research Review*, 2(2): 130-144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Kane, M.T. (2006). *Validation.* In R.L. Brennan (ed.), *Educational measurement* (4th ed.), 17-64. American Council on Education and Praeger, Westport, CT.
- Mehan, H. (1979). *Learning lessons: Social organization in the classroom.* Harvard University Press, Cambridge, MA.
- Mercer, N. and Littleton, K. (2007). *Dialogue and the development of children's thinking: A sociocultural approach.* Routledge, London.
- Michaels, S., O'Connor, C. and Resnick, L.B. (2008). *Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life.* *Studies in Philosophy and Education*, 27(4): 283-297.
- Nystrand, M., Gamoran, A., Kachur, R. and Prendergast, C. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom.* Teachers College Press, New York.
- O'Connor, M.C. and Michaels, S. (1993). *Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy.* *Anthropology & Education Quarterly*, 24(4): 318-335.
- Pianta, R.C., La Paro, K.M. and Hamre, B.K. (2008). *Classroom Assessment Scoring System (CLASS) manual, Pre-K.* Paul H. Brookes, Baltimore, MD.
- Sinclair, J.McH. and Coulthard, R.M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils.* Oxford University Press, London.