frontiers | Frontiers in Education

TYPE Methods
PUBLISHED 26 September 2025
DOI 10.3389/feduc.2025.1625133

OPEN ACCESS

EDITED BY
Serafina Pastore,
University of Bari Aldo Moro, Italy

REVIEWED BY
Julija Melnikova,
Klaipéda University, Lithuania
Paulo Sérgio Garcia,
Universidade Municipal de São Caetano do Sul, Brazil

*CORRESPONDENCE
Niklas Litzenberger
litzenberger@uni-mainz.de

RECEIVED 08 May 2025
ACCEPTED 29 August 2025
PUBLISHED 26 September 2025

CITATION
Litzenberger N, Pysik A and Wurster S (2025) Decoding dynamics in global assessments of teaching. An introduction to global assessments for State Space Grids. *Front. Educ.* 10:1625133. doi: 10.3389/feduc.2025.1625133

Decoding dynamics in global assessments of teaching. An introduction to global assessments for State Space Grids

Niklas Litzenberger[1]*, Andreas Pysik[1] and Sebastian Wurster[2]

[1]Institute of Physics, Johannes Gutenberg-University, Mainz, Germany, [2]Department of Educational Science, Johannes Gutenberg-University, Mainz, Germany

Most state-of-the-art teaching analysis is based on global (i.e., whole lesson) assessment techniques, which cannot account for dynamic processes that characterize learning-teaching relationships. We aim to overcome these limitations by proposing the Global assessments for State Space Grids (GSSG) method, which evaluates the global assessment in terms of the temporal evolution of key indicators and extends Hollenstein's State Space Grids (SSGs). For fast and accessible interpretation, the GSSG method implements established and new stochastic parameters with corresponding graphical elements that go beyond the assessments of the SSGs. These include parameters to measure global relationships between indicators of teaching quality over time, cluster analysis, error estimates, alignment tests, and parameters of different dynamic behaviors that affect global assessments. We introduce this method using aspects of classroom management as an example.

KEYWORDS

teaching evaluation, dynamic teaching analysis, time-series, State Space Grids, dynamic systems

1 Introduction

Teaching is characterized by complex dynamic interactions (Maskus, 1976). For example, the teacher-student relations of motivational and emotional states are ever-changing, complex, and time-dependent (Graf and Agergaard, 2022; Moeller et al., 2022; Pennings and Hollenstein, 2019). However, the influence of this time-dependent dynamic on the research's results is often undocumented in the researcher's method design (Bell et al., 2019) and is not the topic under study (Praetorius et al., 2020).

Researchers typically analyze teaching mainly through teachers' views, students' views, or rater observations. Most studies examine teachers' views by conducting interviews or questionnaires (e.g., Li et al., 2009; Cerbin and Kopp, 2006). Other researchers analyze student assessments through tests, questionnaires, or interviews (e.g., Belova and Eilks, 2015; Mainhard et al., 2011). A third option is direct observation in the classroom or analyzing videotaped lessons or observations made by people in the same room. In this case, trained raters observe and assess specific aspects of teaching (e.g., Welsch and Devlin, 2007; Vrikki et al., 2017). Research designs also consist of combinations of video analysis and teacher and student assessments (e.g., Hattie, 2003; Charalambous and Praetorius, 2018; Ito, 2019; Dorfner et al., 2018). Large-scale studies like PISA or TIMSS also combine assessments of students and teachers by using questionnaires or tests for both (e.g., Mullis et al., 2020; Schleicher, 2019).

However, most research methods rely on single measures of complex constructs and dynamic interactions (Praetorius et al., 2018). The analyzed data rarely has more than two measured points for one lesson (Praetorius et al., 2020), which makes it impossible to analyze time-dependent relationships during the teaching process. For example, responses to questionnaires about teachers' supportive behavior represent evaluations of students only for the entire duration of a class or another defined period, like school years (e.g., Mainhard et al., 2011). In this case, the extent to which the teacher's supportive behavior varies during the lesson cannot be estimated. Since a questionnaire item only represents the average assessment of all time points and of all students, an estimation of the variation during the class cannot be given. Only the distribution of ratings within the class can be analyzed.

An exception is the TIMMS video study, where researchers could analyze videotaped lessons of different years and countries (Martin et al., 2020) and compare sections of lessons (Stigler and Ronald, 2000) or changes in teaching between different years (Burroughs et al., 2019a). However, these comparisons usually average over all lessons, because the goal of the study is to compare teaching at country or year level, and do not delve deeper into the specific time evolution of each lesson's section (Burroughs et al., 2019b).

Due to the lack of research and theory including time-dependent dynamic behaviors, many research groups formulated a call to expand existing theories and develop methods to capture dynamic behavior (Dirk and Nett, 2022; Dietrich et al., 2022; Pekrun and Marsh, 2022; Moeller et al., 2022). We propose a potential solution by developing the concept of State Space Grids further by incorporating methods for global assessments which take development and variation over time into account.

These State Space Grids are based on the time-sampling method established in educational research for some time. The time-sampling method consists of rating indicators mainly through ratings of videotaped lessons and calculating their mean values (e.g., Seidel, 2005; Heinze and Erhard, 2006; Espin and Yell, 1994). In research on teaching indicators are typically predetermined criteria or measures that help identify and quantify specific features or behaviors of the observed lesson (Smith, 1988; Shavelson et al., 1990; Blank, 1993; Dilshad and Iqbal, 2010; Lotz et al., 2013). For example, Heinze and Erhard (2006) analyzed how much time a teacher gives students to answer a question during lessons using this method. They measured the time between the asked question and its answer and calculated a mean value of all measured time intervals. In this case, the time the teacher gives to think about the question would be the measured indicator. Through this method, they gain many different measures each time the teacher waits for an answer in the same lesson. By calculating the mean value of these times measures, they gain an insight into how long the teacher paused for a response on average throughout the lesson. This way, they can consider variations in the teacher's behavior. Time-sampling indicators are a promising approach that can be developed further by analyzing the time-dependent dynamic behavior of the time-sampled indicator to go beyond calculating the average of the time-sampled indicator.

State Space Grids itself originated from research on dynamic interactions (Lewis et al., 1999), was developed by Granic and Lamey (2002) as well as (Hollenstein, 2007), and refined afterward (Hollenstein, 2013). This led to a new research method called State Space Grids (SSG) to give further insight into the dynamic interaction of time-sampled indicators. The method exploits the potential of time-sampled indicators by displaying two indicators in a coordinate system at every sampled time slot (Lewis et al., 1999). This way, we can show and analyze both indicators' dynamic behavior and interactions simultaneously (Hollenstein, 2007) to give further insights into their dynamic interactions (Hollenstein, 2013).

So far, educational research groups use the SSG method mainly to analyze teacher-student interactions (Mainhard et al., 2012; Smit, 2016; Pennings and Mainhard, 2016; Turner and Christensen, 2020; Pennings and Hollenstein, 2019; Scherzinger et al., 2020). For example Scherzinger et al. (2020) investigated how the variability of teacher-student interactions influence teachers' and students' perceptions of their interactions, using the SSG method. Additionally, they compared their results with a questionnaire to analyze the differences in teacher and student views. Based on the SSG method they found that the teacher's view of the student-teacher relationship depended on the variability of their interactions. Psychological researchers also use the method to analyze the influence of innovative moments leading to successful psychotherapy (Ribeiro et al., 2011; Bento et al., 2014), mother-infant interactions (Provenzi et al., 2015), counselor-client interactions (Hoekstra et al., 2023), or coach-athlete interactions in sports teams (Erickson, 2009; Erickson et al., 2011; Meinecke et al., 2019).

Many research groups are already demanding an expansion of existing research methods (Dirk and Nett, 2022; Dietrich et al., 2022; Pekrun and Marsh, 2022). Led by the idea of time-sampled indicators, Moeller et al. (2022) propose the DYNAMICS framework with moment-to-moment ratings to analyze dynamic aspects of teaching. They suggest comparing indicators' ratings across different time points within their framework. With SSG, it is possible to compare indicators across different time points, but we cannot capture the influence of dynamic time developments on global assessments. We need to expand existing research methods to capture this influence of complex dynamics of indicators on global assessments. Hence, we propose the "Global assessments for State Space Grids" method (short: GSSG) by expanding the State Space Grids method.

2 State Space Grids

First, we briefly introduce the State Space Grids (SSG) method because this is the basis of our GSSG method. This SSG method typically involves rating two indicators several times in fixed time intervals (Hollenstein, 2007).

This paper uses a simple example of two indicators to make the method more concrete. Additionally, for more clarity, we use synthetic sample data to analyze the dynamic interactions of the speech distribution of teachers and students. The first example indicator, "verbal engagement students," measures the proportion of time the students speak in a defined period of time. In this example, we use a time interval of 60 seconds. If the students speak

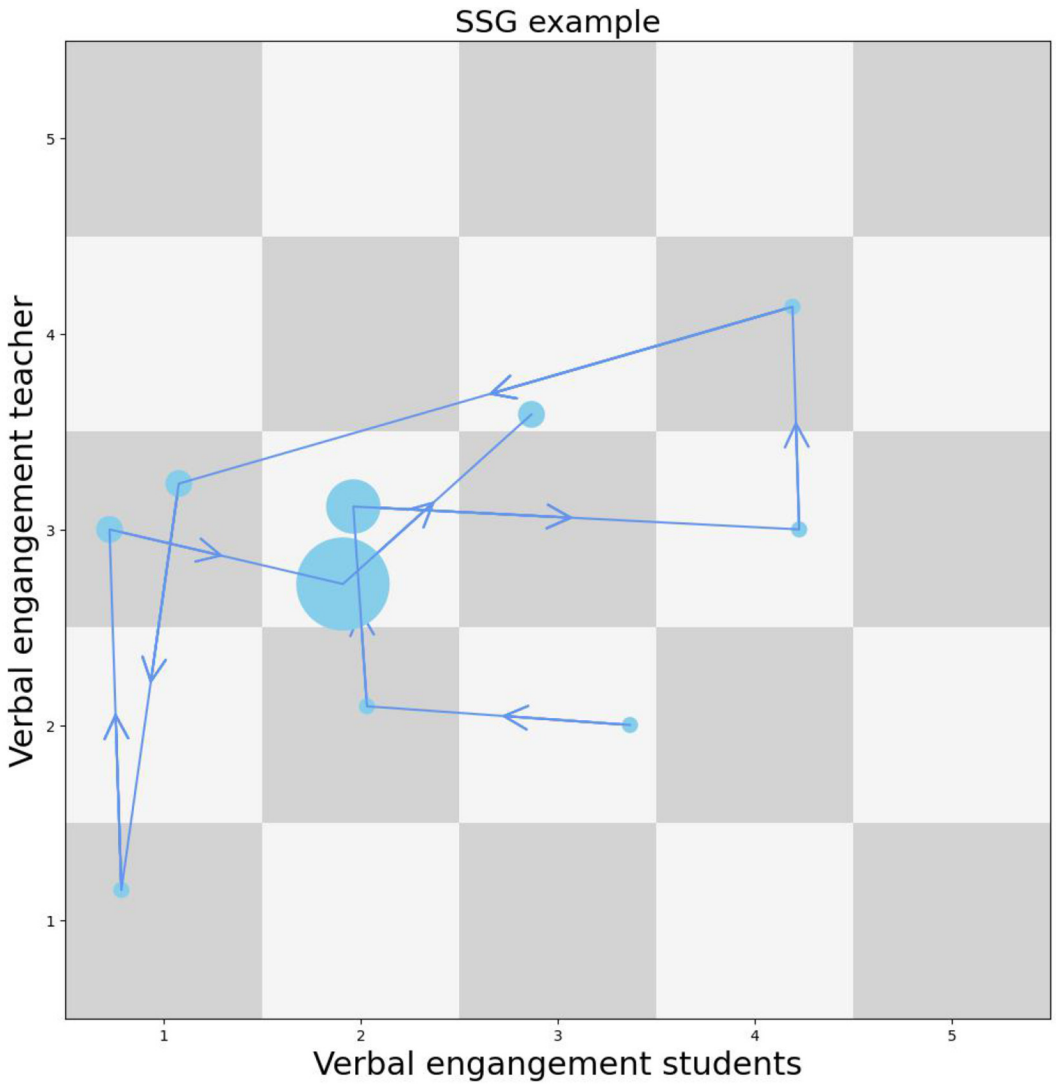

FIGURE 1
Example of a State Space Grid (SSG).

80%–100% of the time, the indicator gets rated as "5." We rate the indicator as "4" if the students speak 60%–80%. Likewise, we award a "3" with 40%–60%, a "2" with 20%–40%, and a "1" with 0%–20%. The second indicator is similar and measures the time the teacher speaks within the 60-second time frame.

In general, it is also possible to gain data for the SSG method through event sampling, where we count the time of a specific event, for example, the friendliness of the teacher toward the students (Scherzinger et al., 2020). However, we will use fixed time intervals as an example because they are easier to follow.

The SSG method uses a coordination system of cells, as in Figure 1, to visualize the dynamic behavior of the indicators. We will first look at a first 60-second time frame: Suppose the teacher greets the students, the students answer shortly, and after a few seconds of silence, one student starts to ask a question in the remaining time of the first minute. In that case, the first rating could be (3, 2). The exact position of the rating within the square of (3, 2) is irrelevant and serves the purpose of nearly no overlap of many ratings within the same cell.

In the second minute, the students speak as much as the teacher, and the rest of the time, the room is silent. This situation would lead to a rating of (2, 2). Afterward, in the third minute, the teacher and the students start a conversation where the teacher speaks more, leading to a (2, 3) rating. This situation continues for a few minutes, leading to the same (2, 3) rating for several minutes. The SSG method indicates this by increasing the circle's radius in the (2, 3) cell. A line connects these ratings with an arrow to visualize that time is moving forward. Should the same rating appear again after changing, a new circle appears in the corresponding cell, like in the (2, 3) cell. This way, the State Space Grid method simultaneously visualizes both indicators' dynamic processes.

At first sight, most ratings remain at and near (2, 3) since their circle's diameters are larger than in other ratings. For the interpretation, this means that the teacher mainly speaks a little more during the lesson than the students. Such a location in the grid is called an attractor (Lewis et al., 1999). In contrast, ratings that are rarely visited [i.e. (1, 1)] or not visited at all [i.e. (5, 1)] are called repellors (Hollenstein, 2007). For example, a repellor exists

when the teacher and the students speak most of the time and simultaneously, in the rating (4, 4).

Hollenstein (2013) mainly looks at point attractors, specific states/cells in the grid where most ratings are "pulled." In the case of Figure 1, the cell (2, 3) is such a state that it is a point attractor. The SSG method specializes in analyzing specific states and transitions between different cells. The SSG method analyzes the ratings by using the visualization of the SSG plots, as well as parameters for cell assessments and the whole grid.

These whole-grid parameters range from the mean duration, which measures how long ratings occupy the same cell on average, to a measure called Dispersion, which compares the overall duration in each cell with the overall duration of all events. These parameters can describe how long the ratings visit the same cell overall. They indicate if the ratings are spread evenly along all cells or only in a few specific cells, like in cell (2, 3) in Figure 1. In our example, the mean duration of the cell (2, 3) is high, and the whole-grid mean duration and the Dispersion are low, indicating that the teacher mainly speaks more than the students. However, one must know that the cell (2, 3) is likely an attractor to measure this effect.

In this case, cell assessments are helpful that only look at one specific cell. For example, the mean return time measures the average time needed for a rating to return to a specific state or the mean duration. This parameter indicates whether a specific cell can be an attractor, which is why the measures of the SSG method help verify the impression that (2, 3) is an attractor (Turner and Christensen, 2020).

The SSG measures also include two entropy assessments that indicate if there is a specific pattern of jumps between two or more cells that happens often. One example is when only the teacher speaks at 1-min intervals and then only the pupils in the following interval. In this case, the SSG would have one big jump between (1, 5) and (5, 1), and the entropy assessments can reflect that and measure how dominant the patterns are. For example, Figure 2 has two patterns: the big jump between the cell (5, 1) and (1, 5) and jumps between (2, 4) and (1, 5). A higher entropy indicates, in this example, that there are some patterns of changes in the speech distributions.

The SSG method can further measure transitions between specific cells with these cell assessments. Two assessments can analyze transitions between two cells, which indicates how often a transition happens in comparison to its duration within one cell and whether there is a specific pattern. Through these two assessments, the SSG method tells us that the transition between (5, 1) and (1, 5) is not as dominant as between (1, 5) and (2, 4) because the number of transitions and the duration in both states are lower for the transition between (5, 1) and (1, 5).

However, it is apparent that in Figure 2 are two clusters: one in the upper left area and one in the lower right area. In such a case, only looking at specific cells is not sufficient. The SSG method solves this problem by collapsing the clusters into one cell. This way, we can look at the upper left and lower right clusters each as one cell and measure the transition between these collapsed cells. In this case, the SSG method looks at these two collapsed cells like two point attractors and can also apply its cell assessments, like the mean return time.

We see more potential in analyzing such clusters and patterns. As Hollenstein says, "the scope and power of State Space Grids as a dynamic analysis tool has not yet been fully realized and it is up to the current and next generation of researchers to push and extend the technique into the greater prominence it deserves" (Hollenstein, 2012, p.15). Since one "method of attractor analysis that has yet to be tried is to examine attractor change and stability over time on the state space" (Hollenstein, 2012, p. 77), we will and can look at these clusters or global attractors over time. Compared to Hollenstein's point attractors, we look at global attractors, clusters of ratings beyond just one cell, like the upper left and lower right clusters in Figure 2 to compare their time developments and influences on global assessments. We propose a method that defines and looks further into these global attractors and finds more global patterns than just reoccurring patterns of cell transitions.

Since we mainly extend the State Space Grid method by looking at dynamic effects on global assessments, we call this method the Global assessments for State Space Grid method, or the GSSG method.

3 Parameters for global assessments for State Space Grids

In this chapter, we will introduce step by step the new parameters of our method to delve deeper into global assessments for dynamics and go beyond the assessments of State Space Grids. We will not limit the number of dimensions in our method to two, but will allow for the extension of the method to any number of dimensions. This way, we can, for example, look at the speech distribution of the teacher and all students and simultaneously at the changes in the speech distribution of every single student and the teacher. Another possibility is to extend the two-dimensional space with two other entirely different indicators, like an indicator for the distribution of time when the students listen to the teacher and an indicator for the distribution of time when the teacher listens to the students. However, throughout this paper, we will mainly illustrate the implications of the GSSG method as an example of two indicators for speech distribution. We will start by formulating global assessments for variability (i.e., the change in the ratings) and then look at specific dynamic patterns and their global assessments.

3.1 Measuring variability with global assessments

First, we will examine to what extent a single global estimation can represent the whole duration of the rated instance. In the GSSG method, we understand a global estimation for the ratings as an indication of where most of the ratings are with respect to the overall values of all ratings. In the case of Figure 1, a global estimation is the cluster's center around (2, 3). The most straightforward way to gain such a global rating estimation is through a two-dimensional mean value with a traditional mean value for every axis. This approach differs from a point attractor,

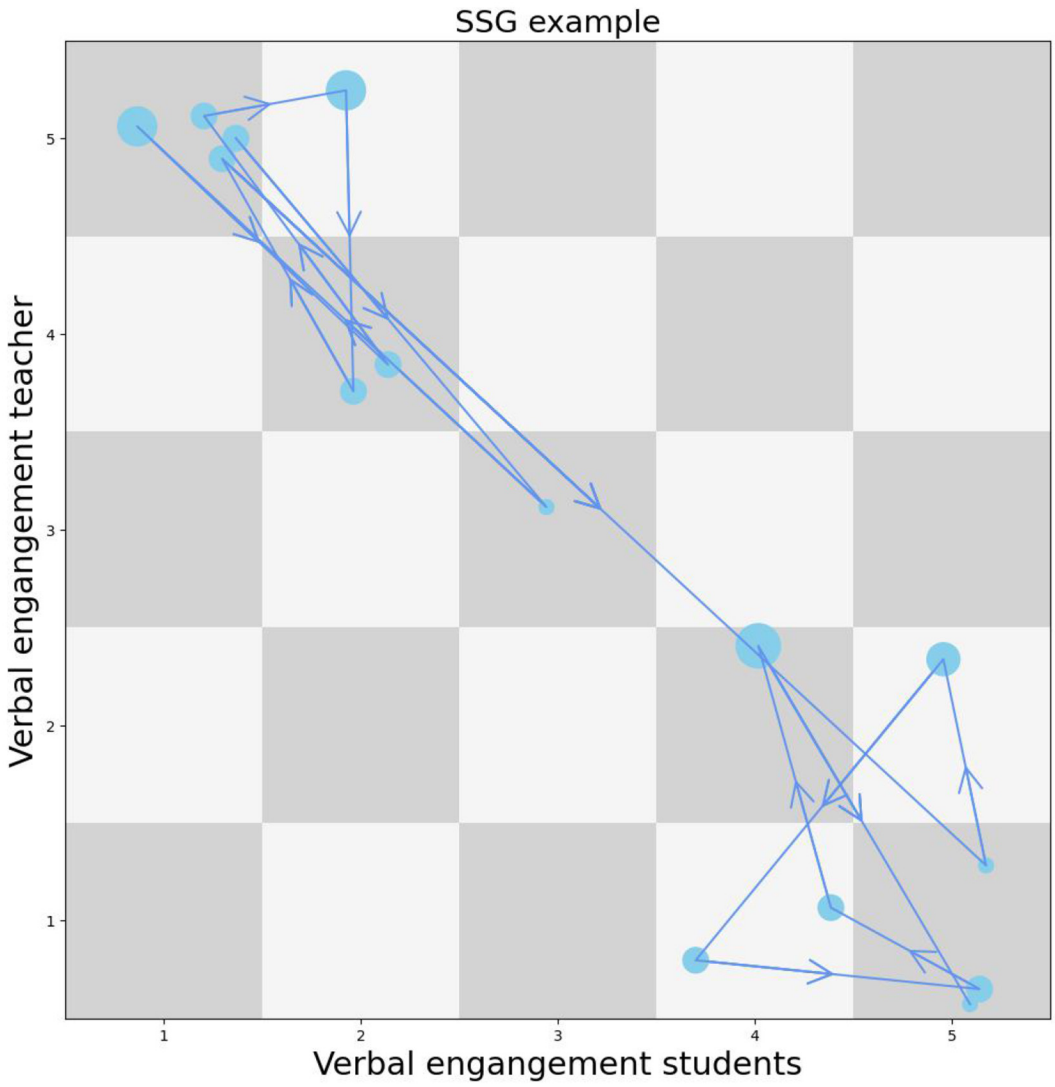

FIGURE 2
Example of a State Space Grid (SSG) with two phase transitions.

which only describes single cells, since we can describe a single cluster by considering the whole duration and all ratings. Hence, we call such a location in the grid a global attractor. In the case of Figure 1, the position of a global attractor is in (2.05, 2.97), which indicates that there is nearly no shift toward more or toward less speaking time to the teacher or the students. If there is more than one global attractor, like in Figure 2, we can look at each global attractor separately. In the next chapter, we look closer at the case of several global attractors. This chapter will first focus on the case of a single global attractor.

A mean value for a global estimation by itself is not a novel idea since it is a typical way to gain an estimation of a set of ratings. What makes the GSSG method interesting is that we can describe how strong this global attractor is and, thus, to what extent it can describe a global behavior. As Hollenstein suggested, to "really establish a comprehensive sense of the attractor landscape revealed by the trajectory patterns on the State Space Grid, it may be best to use a combination of methods and variables" (Hollenstein, 2012, p. 77). This is why we will introduce different parameters to measure how and to what extent the global attractor pulls the ratings near its location in the grid.

The first aspect we will examine is to what extent the ratings change in time and, if they change, how much. To measure this change in the ratings, we calculate the distance each rating $x_{i,j}$ of the time instance i and the indicator j while accounting for the total number of ratings n and indicators d. This calculation leads us to a measurement of the average distance the ratings travel each time step. Thus, we call this parameter the Mean Travel Distance (short: *MTD*):

$$MTD = \frac{1}{\sqrt{d}(n-1)} \sum_{j=1}^{n-1} \left| (x_{1,j}, \ldots x_{d,j}) - (x_{1,j+1}, \ldots, x_{d,j+1}) \right|. \quad (1)$$

The Mean Travel Distance comes from an idea by Hollenstein that we can calculate each rating transition as a distance. However, compared to Hollenstein's idea, we divide this distance by the

number of jumps $(n-1)$ in the grid and the average increase in the dimensional distance for higher dimensions $\sqrt{d}$. This way, we do not just add up all jump distances but gain an average of all jumps, which we can compare to other GSSGs with a different number of ratings and indicators. Thus, the Mean Travel Distance calculates the average distance a rating jumps in each dimension in the grid. This parameter is also a more precise measurement of the overall duration in each cell than the Dispersion, which only looks at the overall duration in each cell and can not take the frequency and distance of each transition into account.

Looking at our example in Figure 1, the Mean Travel Distance is 0.28, meaning that the average speech distribution changes by 0.28. This value of 0.28 equals an average change of 5.6% in the speech distribution of the teacher and the students. In this case, the highest possible value of the Travel Distance is 4, which would be the case where the ratings jump every time step from 1 to 5 and back again. The lowest possible value for the Mean Travel Distance is always 0, which is the case where the ratings do not change at all and thus have a jump distance of 0. Thus, a value of 0.28 is relatively low, which means that the global attractor of (2.05, 2.97) is not influenced much by the frequent changes in the ratings.

In contrast, a high *MTD*, like 0.8, would indicate that the speech distribution changes frequently, and when it changes, it changes a lot. In such a case, a single global estimation can not describe the speech distribution. Instead, we can describe the overall speech distribution as constantly changing, which is also a meaningful and comparable effect by itself.

The Mean Travel Distance by itself is not a sufficient parameter to describe the influence of the changes in ratings on the global attractor alone. For example, there can be a case where each time a rating changes, it keeps increasing. In such a case, like Figure 3, the Mean Travel Distance is low with 0.16, but there is no global attractor where most ratings are, like in Figure 1. We will take a closer look at this case later because this is a particular pattern we will analyze. For now, we will solve the problem of the mean travel distance being insufficient to indicate if there is a global attractor that can represent a global estimation.

A traditional way to measure if a mean value can represent the total data is the standard deviation and the standard error of a mean (Barde and Barde, 2012). However, we have more than just one indicator, and we want to analyze how well the global attractor can represent all indicators in the grid. Thus, we average the standard deviations to measure the overall deviation from the global attractor and normalize the result so that the maximum value is 1. We call this parameter the Mean Standard Deviation (short *MSD*), which we calculate through:

$$MSD = \frac{1}{d}\sum_{i=1}^{d}\frac{2}{o_i}\sqrt{\frac{1}{n}\sum_{j=1}^{n}(\overline{x_i}-x_{i,j})^2} \tag{2}$$

where o_i is the highest value of each indicator.

For a 5×5 grid with a rating system between 1 and $o_1 = o_2 = 5$ on two indicators, like in all of our examples, this parameter ranges from 0 when all ratings are in one cell and 1 when there is an equal amount of ratings in $(1, 5)$ and $(5, 1)$. Through this parameter, we can now differ between the case of Figure 3 with an *MSD* of 0.47 and Figure 1 with an *MSD* of 0.291. Thus, the *MSD* and the *MTD* can tell us if there is a global attractor where the frequency and distance changes in the ratings are low (*MTD* low) and the rating remains in a similar location (*MSD* low). For our example, a low *MSD* and low *MTD* mean that the overall speech distribution indeed can be described through the (2.05, 2.97) rating. Whenever the students or the teacher speak less or more than that, they keep returning to the same distribution (low *MSD*), return fast to this state, and stay there for a long time (low *MTD*).

In comparison, in Figure 3, the global rating is not meaningful (*MSD* high), because the teacher and the students change their speech distribution too much. However, it rarely happens that the teacher or the students change how they speak, but when they speak more, they remain in this state for longer (*MTD* low). This low *MTD* and high *MSD* indicate that the teacher and the students go through different but long remaining states where there must be a cause for why they speak more and more, like a heated debate. Through these two parameters, we can now prove, with calculated values, that such states exist, and we can compare this effect size with, for example, different classes.

Now, we can calculate the extent to which a global estimation is meaningful, and we know why it is meaningful (low *MTD* and *MSD*) and why it is not (high *MTD* or *MSD*). Furthermore, we also want to know if only a single indicator is why we can not represent all ratings through a single estimation or if all indicators contribute equally to this cause. We calculate this effect through the absolute differences in the distance changes and the standard deviations. This approach leads us to the Standard Deviation Difference (short *SDD*):

$$SDD = \frac{2}{d(d-1)}\sum_{i=1}^{d}\sum_{k=1}^{i}\left|\frac{2}{o_i}SD_i - \frac{2}{o_k}SD_k\right|. \tag{3}$$

and the Travel Distance Difference (short TDD):

$$TDD = \frac{2}{d(d-1)}\sum_{i=1}^{d}\sum_{k=1}^{i}|TD_i - TD_k| \tag{4}$$

where TD_i is the one-dimensional Travel Distance of the ith indicator and SD_i its Standard Deviation. We obtain the Travel Distance by calculating the *MTD* for a single indicator. The additional factors ensure that the values of *SDD* and *TDD* range between 0 and the maximal value of the corresponding parameter.

In the case of our example in Figure 1, the Standard Deviation Difference is 0.06, indicating that one party (in this case, the students) is very slightly but not significantly represented less in the global estimation than the other party. A Travel Distance Difference of 0.000 indicates that both the teacher and the students change the amount they speak equally frequently. These low values indicate that both parties are equally well represented in the global estimation, and the dynamic changes of both are similar ($TDD \approx SDD \approx 0$).

The analysis of dynamic changes with global assessments now depends on four parameters: two for the average changes (*MTD* and *MSD*) and the dynamic differences between each indicator (*SDD* and *TDD*). Analyzing these four parameters can be time-consuming, especially with many lessons and plots. In the next step, we will make the analysis of these four parameters easier and faster.

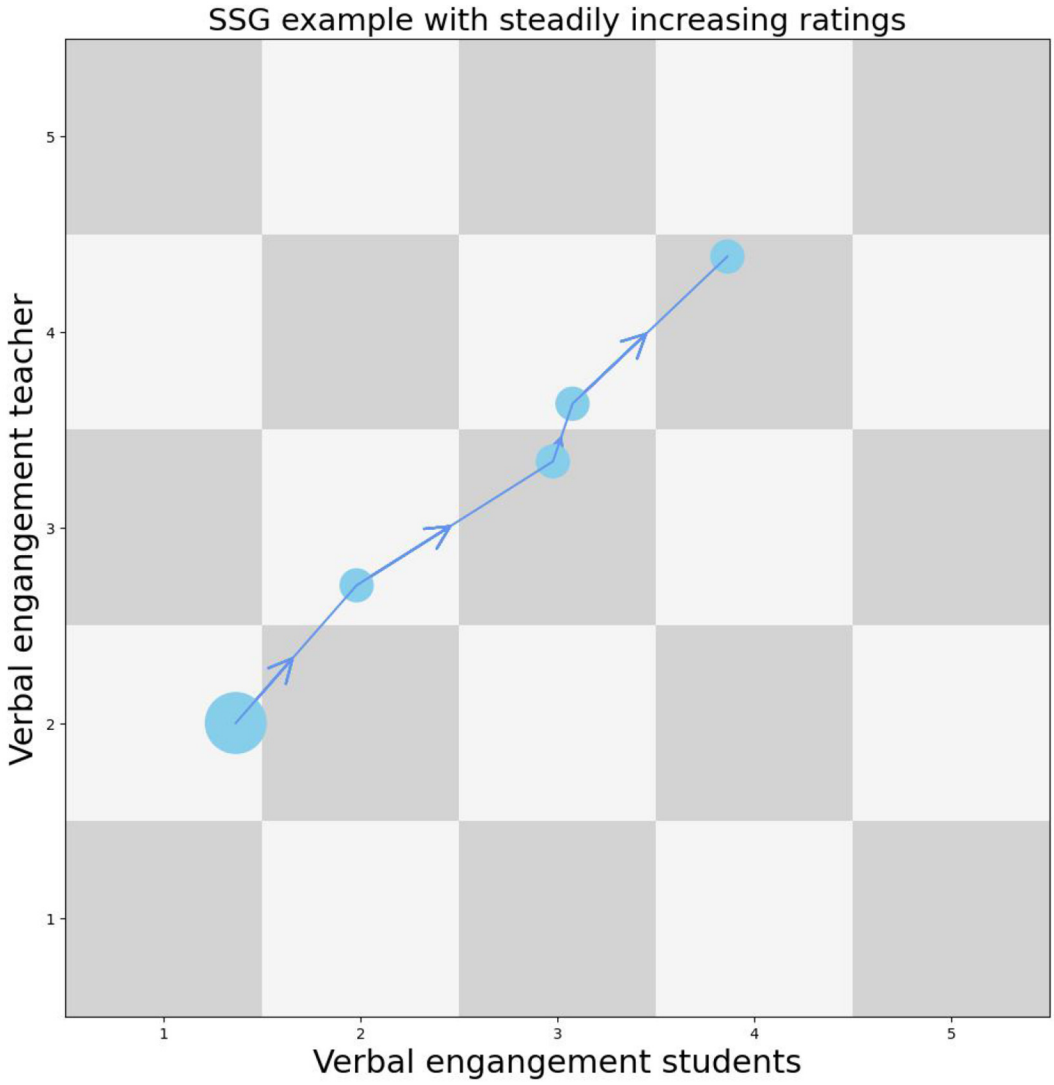

FIGURE 3
Example of a State Space Grid (SSG) steadily increasing ratings.

A particular strength of the State Space Grid method is its two-dimensional graphical illustration of the dynamic changes of the indicators in time through the ratings represented on the grid. To analyze global attractors and the influence of dynamic changes on them at first glance, we will lean on the strengths of the SSG's graphical representation by adding more elements to the grid. We visualize global attractors through a red circle in the grid on the position of its values on the x- and y-axis. To illustrate each indicator's Travel Distance, we draw a black ellipse around the global attractor with the Travel Distance of the x-axis as the semi-major axis and the Travel Distance of the y-axis as the semi-minor axis of the ellipse. Additionally, we draw a second red ellipse for the Standard Deviations for each axis.

Looking at Figure 4, we can find the global attractor in $(2.05, 2.97)$. In contrast to the blue circles for the ratings, which are at random positions in their corresponding cell, the position of the red dot, the width, and the height of both ellipses are exact representations of the corresponding parameter. The small size of the black ellipse indicates that the Mean Travel Distance is low, and since the ellipse is approximately circular, we also know that the Travel Distance Difference is zero. In comparison, the red ellipse has a slightly higher width than height, which tells us that the Standard Deviation Difference is small but higher than zero and that the Standard Deviation for the verbal engagement of the students is higher than for the teacher. Overall, we can see at first glance that the Mean Travel Distance and the Mean Standard Deviation are similar (ellipses have a similar size), that the global attractor can represent the whole duration (both ellipses are small), and that both indicators contribute equally to the dynamics of the lesson (both ellipses are circle-like).

In comparison to Figures 4, 5 shows a lesson where the amount the teacher speaks changes frequently ($MTD = 0.707$), but the students keep speaking the same amount ($TDD = 1$). However, overall, most ratings are near the cell $(2, 3)$ (the red ellipse is flat), which means that the state where the teacher speaks slightly more than the students often occurs ($MSD = 0.262$). In this case, a single

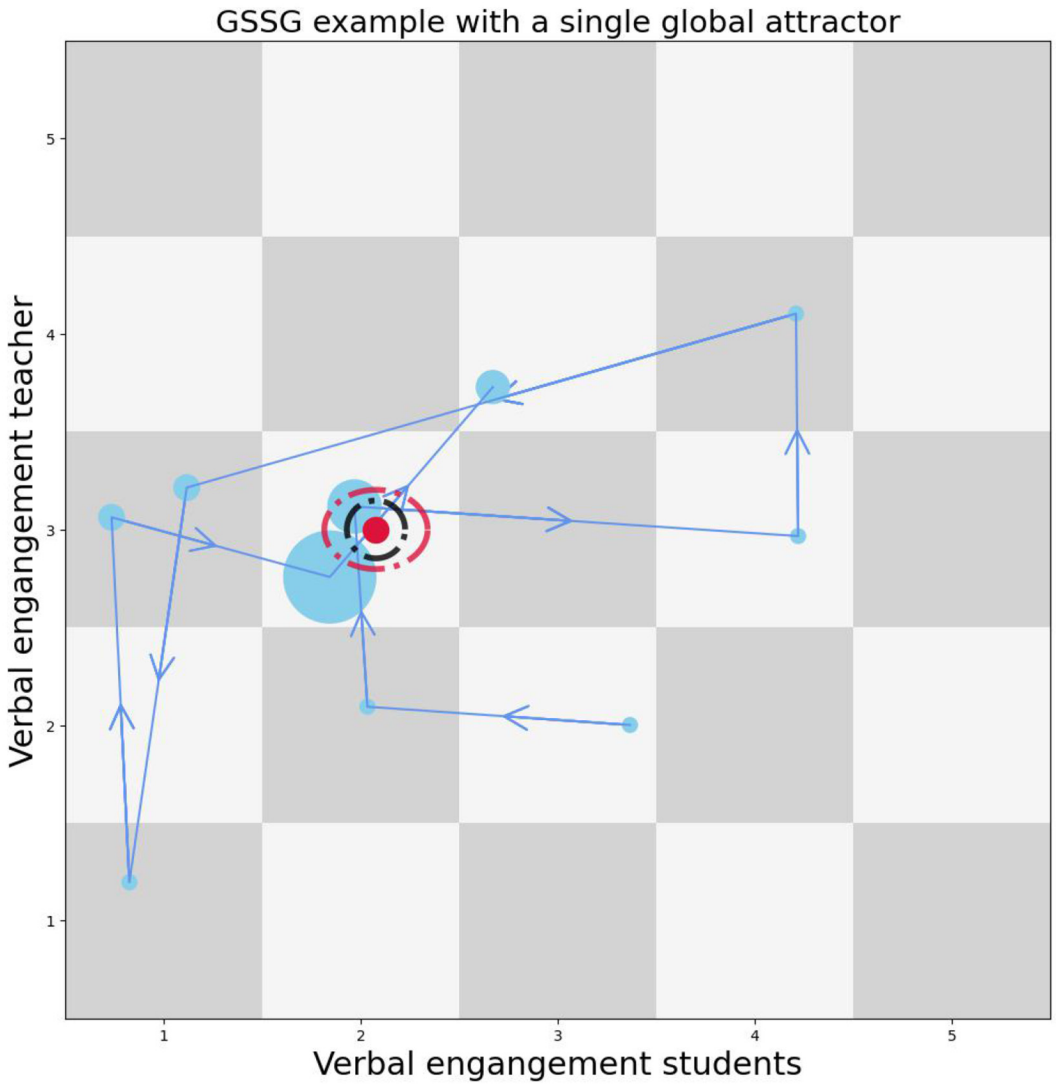

FIGURE 4
Example of a GSSG with *MTD* = 0.271, *TDD* = 0.000, *MSD* = 0.291, *SDD* = 0.06.

global estimation is not meaningful because the speech distribution changes too often. Furthermore, at first glance, we can see that the cause of this dynamic change is the teacher and not the students (the yellow ellipse stretches out on the axis of the teacher).

Next, we will look at five specific patterns and corresponding global assessments, which we can measure using the GSSG method.

3.2 Measuring dynamic patterns with global assessments

This chapter will look at the grid differently by analyzing dynamic patterns in the whole rating duration and formulating global assessments. We will look at alignment patterns, overall increases or decreases, and similarities between indicators.

The first type of pattern we look at is alignment. Linear alignments like in Figure 3 are very common in scientific research and are usually described through a fit and a parameter for the goodness of the fit. We will also implement this traditional approach in the GSSG method with two types of alignments.

The first type of alignment is the linear alignment of Figure 3. For the speech distribution, a sufficient linear fit with a positive slope means that whenever the teacher speaks more, the students speak more simultaneously, like in Figure 3. The slope of 0.71 also tells us that a change in the teacher's speaking amount changes the students' amount more than the students' speaking amount changes the teacher. In contrast, a negative slope indicates that the students speak less whenever the teacher speaks more, like in Figure 2.

The second type of alignments we look at are grids, where whenever one indicator changes, the other indicator remains the same. We use vertical and horizontal fits for these alignments. An example of such a pattern is Figure 5. In terms of the example, such a pattern indicates that no matter how much the teacher speaks, the students barely speak.

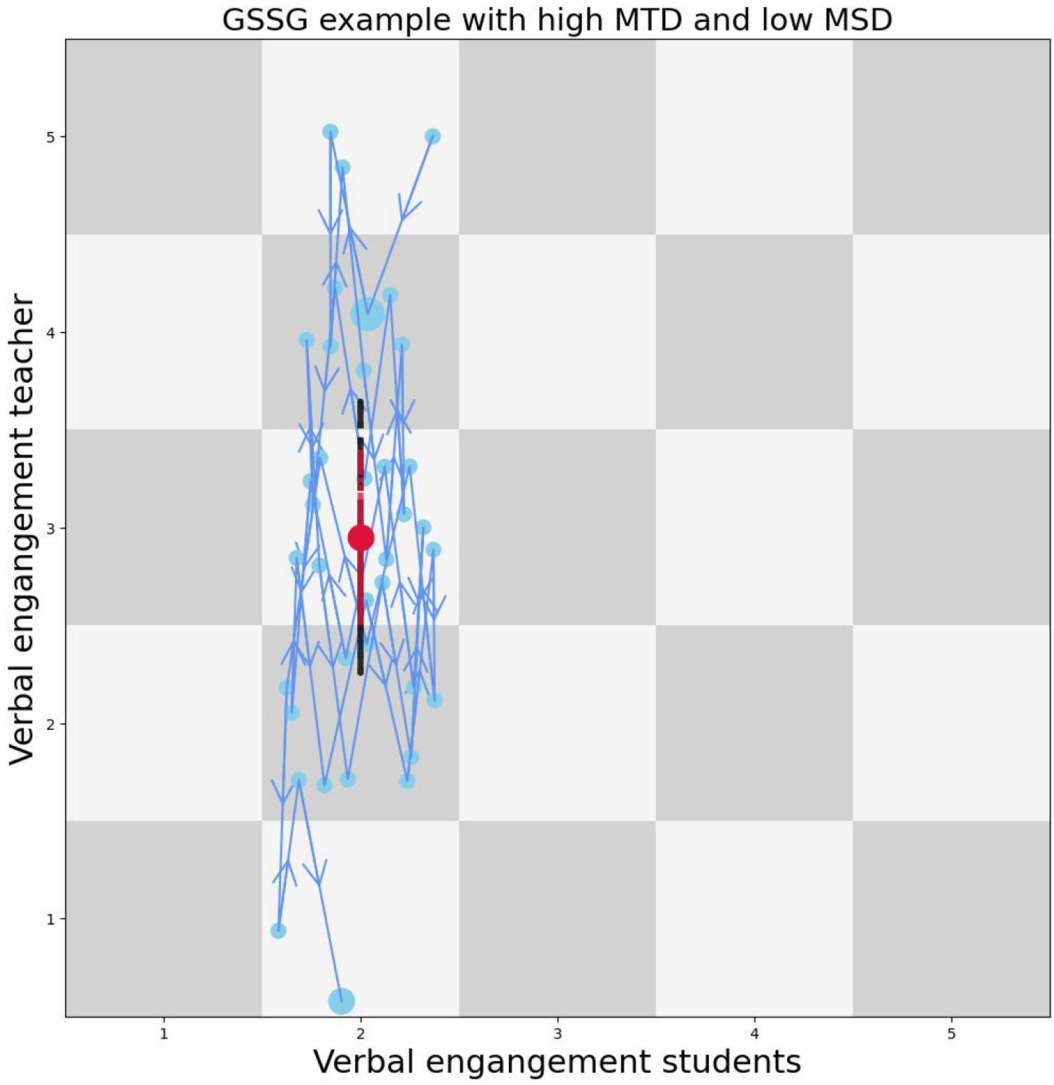

FIGURE 5
Example of a GSSG with $MTD = 0.707$, $TDD = 1$, $MSD = 0.262$, $SDD = 0.420$.

We measure the goodness of these fits through a R^2 test. The R^2 test is a conventional test that compares the ability to describe data through an inserted fit and a mean value. For this test, an acceptable R^2 value is higher than 0.1 in social science research (Ozili, 2023). In the case of Figures 3, 5 a R^2 of 0.87 and 1 indicate that a single global estimation can not represent the whole duration. We know that through the Mean Standard Deviation and the Mean Travel Distance. However, now we also know that we can use the slope of $m = 0.71$ for Figure 3 and the fixed value $x = 2.00$ with a constantly changing y-axis for Figure 5 instead for a global representation. However, we can also calculate one more pattern in these two examples.

This second type of pattern we look at is global trends in time. For example, suppose the amount of spoken time increases on the students' side during a lesson. In that case, a single global estimation for the speech distribution can be meaningful, but with the additional information, we know one indicator steadily increases. We also know that the representation through the global estimation for the first half is lower than the second half.

We calculate this rating increase by measuring if a rating jumps to a higher or a lower next rating. If it goes to a higher rating, we add a $+1$. If it goes to a lower rating, we add a -1. A mathematical way to gain a $+1$ or -1 is the indicator function $\mathbf{1}_A(j)$ which results in 1 if the inserted value j (in our case, a step in time) is part of a specified set A (in our case a set for all time steps with an increasing rating). If j (the time step) is not part of the set A (meaning it has no increasing rating), the indicator function returns a zero. We repeat this step of adding $+1$ or -1 for all $n-1$ jumps for all n ratings and all d indicators. Then, as a final step, we normalize the sum result to gain a final value between -1 (all ratings decrease in time) and $+1$ (all ratings increase in time). This idea leads us to a new parameter we call the Mean Travel Tendency (short: MTT), which we calculate through:

$$MTT = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{o_i - 1}\sum_{j=1}^{n-1}\left(\mathbf{1}_{\{j\in\mathbb{N}|x_{i,j+1}-x_{i,j}>0\}}(j) - \mathbf{1}_{\{j\in\mathbb{N}|x_{i,j+1}-x_{i,j}<0\}}(j)\right). \quad (5)$$

Looking back at Figure 3, we already know that there exist states where the speech distribution remains the same and then changes and remains in the new state for a long time again (*TDD* low). We can also show through the Mean Travel Tendency of 0.625 that it increases when the speech distribution changes. In comparison, in Figure 5, the ratings are decreasing with a Mean Travel Tendency of −0.50.

Similar to the Mean Travel Distance and the Mean Standard Deviation, we can also answer whether all indicators drive this increase or decrease in time or if just a few indicators cause the increase/decrease. We calculate this effect through the Travel Tendency Difference (short: *TTD*):

$$TTD = \frac{2}{d(d-1)}\sum_{i=1}^{d}\sum_{k=1}^{i}\left||TT_i| - |TT_k|\right| \quad (6)$$

where TT_i is the one-dimensional travel tendency of the ith indicator.

Using the Travel Tendency Difference, we can compare through measurements the trend for rating increases or decreases, like in Figure 3, where both indicators contribute almost equally to the increase of the speech distribution ($TTD = 0.25$). The trend in Figure 5 is only because of one indicator ($TTD = 1$). Combined with a linear fit for Figure 3, we also know that the teacher keeps speaking more, and the students keep speaking more simultaneously.

As with the variability parameters, we want to implement graphical elements to simplify the analysis of patterns. The graphical element for the alignments are fits for the three cases: linear, vertical, and horizontal. We visualize the goodness of fit through the color of the fit. The better the goodness of fit is, the darker and greener the fit gets. Thus, the linear fit for Figure 6 describes the ratings (the color is dark green and $R^2 = 0.87$), and the vertical and horizontal fits do not describe the ratings (the color is white and $R^2 = 0$).

We color each rating in a darker blue to implement a graphical element for the ratings' direction. This way, we can see, at first glance, that the ratings of Figure 6 both increase in time and the ratings for Figure 7 decrease in time on the y-axis.

The last type of pattern we look at is similarities in both ratings. We will measure to what extent both ratings are the same and if there are trends in time, for example, if both ratings are the same in the beginning and different in the end. To measure this effect, we calculate the difference between the ratings $x_{i,j}$ in all time intervals j. Then, we give a weight $\frac{n-j}{n}$ to these differences in a way that differences in the beginning contribute more than differences in the end. We summarize all these weighted differences and normalize them so that the highest value is 1 (all ratings are the same) and the lowest value is 0 (all ratings are different). This idea leads us to the Beginning Similarity (short: *BS*):

$$\mathrm{BS} = 1 - \sqrt{\frac{2}{(n+1)}\left(\sum_{i=1}^{d}\left|o_i - 1 - \frac{o_i - 1}{d}\right|\right)^{-1}\sum_{j=1}^{n}\frac{n-j}{n}\sum_{i=1}^{d}\left|x_{i,j} - \frac{\sum_{i=1}^{d} x_{i,j}}{d}\right|}. \quad (7)$$

A Beginning Similarity near 1 indicates that the ratings, in the beginning, are similar. For our example, this means that the teacher and the students speak the same amount in the beginning, and then either the teacher or the students speak less than the other, like in Figure 8 where the Beginning Similarity is $BS = 0.81$. This high value indicates a behavior change in time that we must consider in a global estimation. In this case, a global attractor can still describe the whole duration, like in Figure 8 (both ellipses are tiny). However, now we can also consider that the students and the teacher speak the same amount in the beginning.

The Beginning Similarity, as its name suggests, can not tell us anything about the similarity in the end. For example, it can also be the case that both ratings are similar at the end and the beginning, like in Figure 6. Likewise, we still can not show through calculations that in Figure 8, the ratings in the end are not similar. Thus, the Beginning Similarity by itself is not sufficient. Thus, we also calculate the similarity in the end with the formula for the Ending Similarity (short: *ES*):

$$\mathrm{ES} = 1 - \sqrt{\frac{2}{(n+1)}\left(\sum_{i=1}^{d}\left|o_i - 1 - \frac{o_i - 1}{d}\right|\right)^{-1}\sum_{j=1}^{n}\frac{j}{n}\sum_{i=1}^{d}\left|x_{i,j} - \frac{\sum_{i=1}^{d} x_{i,j}}{d}\right|}. \quad (8)$$

The only difference between the Ending Similarity and the Beginning Similarity is that the Ending Similarity weights the ratings in the end higher with the factor $\frac{j}{n}$ in the sum.

Through both indicators, we can now calculate to what extent the ratings are similar in the beginning $BS = 0.81$ and different in the end $ES = 0.55$. In contrast, in Figure 6, the ratings are as similar in the beginning as in the end ($ES = 0.62$ and $BS = 0.63$). However, we still can not calculate how much one rating affects the change in similarity.

To gain a measurement for the influence of a single indicator on the similarity of all indicators, we calculate the trend of the differences between two indicators, x_i and x_k, and weigh them over time. This idea leads to a calculation of the trend at the beginning and a calculation of the trend at the end. Thus, we call these parameters the Beginning Similarity Trend (short *BST*):

$$\mathrm{BST}_{x_i,x_k} = \frac{2}{n-1}\frac{1}{\max(o_i,o_k)-1}\sum_{j=1}^{n}\frac{n-j}{n}(x_{i,j} - x_{k,j}). \quad (9)$$

and the Ending Similarity Trend (short *EST*):

$$\mathrm{EST}_{x_i,x_k} = \frac{2}{n-1}\frac{1}{\max(o_i,o_k)-1}\sum_{j=1}^{n}\frac{j}{n}(x_{i,j} - x_{k,j}). \quad (10)$$

A positive value near 1 indicates that if the ratings are not similar, they are higher on indicator x_i. In contrast, a negative value near -1 indicates that if the ratings are not similar, they are higher on indicator x_k.

In terms of our example in Figure 8, an Ending Similarity Trend of $EST = -0.20$ tells us that as soon as the teacher and the students

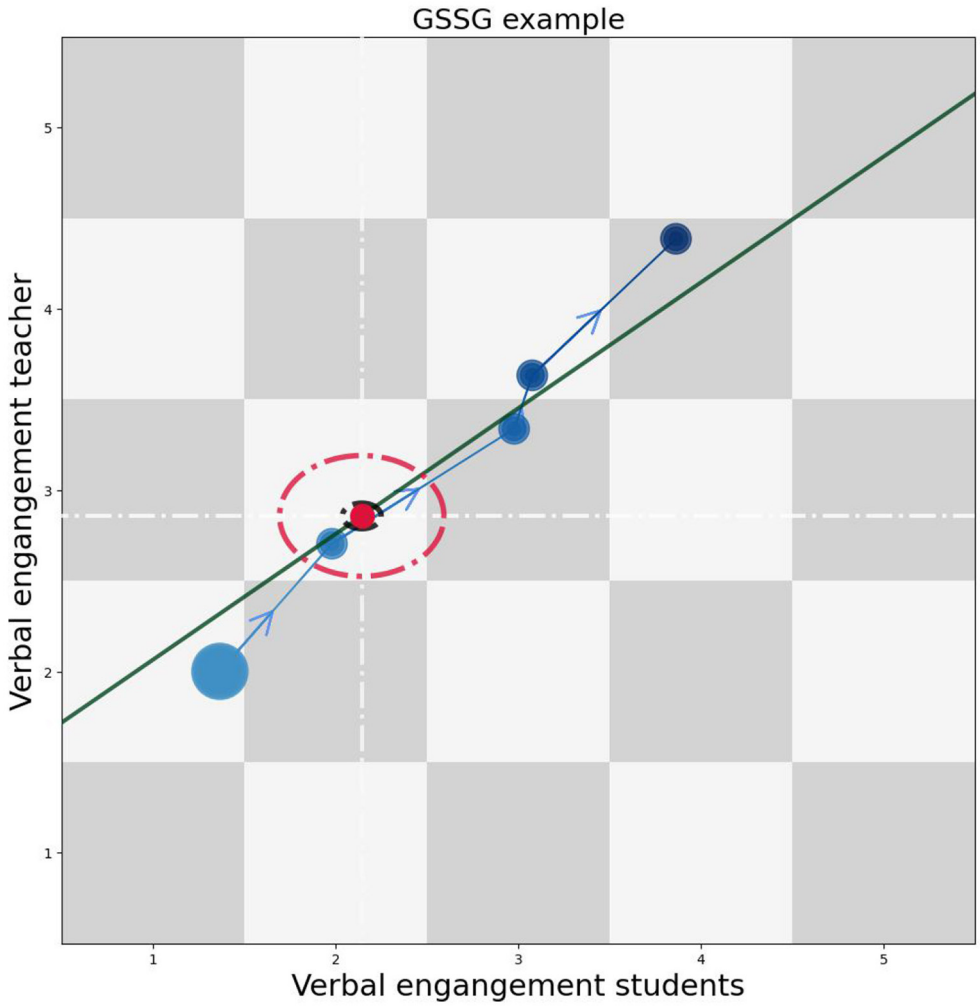

FIGURE 6
Example of a GSSG with $R^2_{linear} = 0.87, R^2_{horizontal} = R^2_{vertical} = 0$, $MTT = 0.625$, and $TTD = 0.25$.

do not speak the same amount, the teacher speaks more in the end, but not for a long time. The Beginning Similarity Trend of $BST = -0.04$ is not engaging in this case because we already know that the ratings are similar in the beginning.

In contrast to the State Space Grid method, we can now also calculate:

- to what extent a global estimation is meaningful (low Mean Travel Distance and low Mean Standard Deviation),
- if a global estimation is not meaningful, to what extent are all indicators the cause (low Travel Distance Difference and low Standard Deviation Difference) or are only a few indicators the cause (high Travel Distance Difference and/or high Standard Deviation Difference),
- to what extent the ratings of all indicators increase or decrease (positive or negative Mean Travel Tendency) or if just a few indicators increase or decrease (high Travel Tendency Difference),
- to what extent do two indicators increase/decrease each other at the same time (high R^2 for a linear fit) or lead to no change in the other indicator (high R^2 for a vertical/horizontal fit),
- are all ratings similar at the beginning (high Beginning Similarity) and/or in the end (high Ending Similarity),
- and if two indicators are not similar, which is higher in the beginning (positive/negative Beginning Similarity Trend) or in the end (positive/negative Ending Similarity Trend)?

Next, we will measure the rater disagreement on these calculations and provide a way to consider their influence separately for each calculation.

3.3 Integrating the influence of rater disagreement on global assessments

We can have multiple raters, especially when analyzing lessons. Researchers often use parameters to determine the overall disagreement of all ratings and raters, like Cohen's Kappa (Cohen, 1960) or interrater correlation coefficients (Goodwin, 2001). However, for these parameters for rater disagreement, the ratings have to be unrelated (Brennan and Prediger, 1981; Fisher, 1938) and

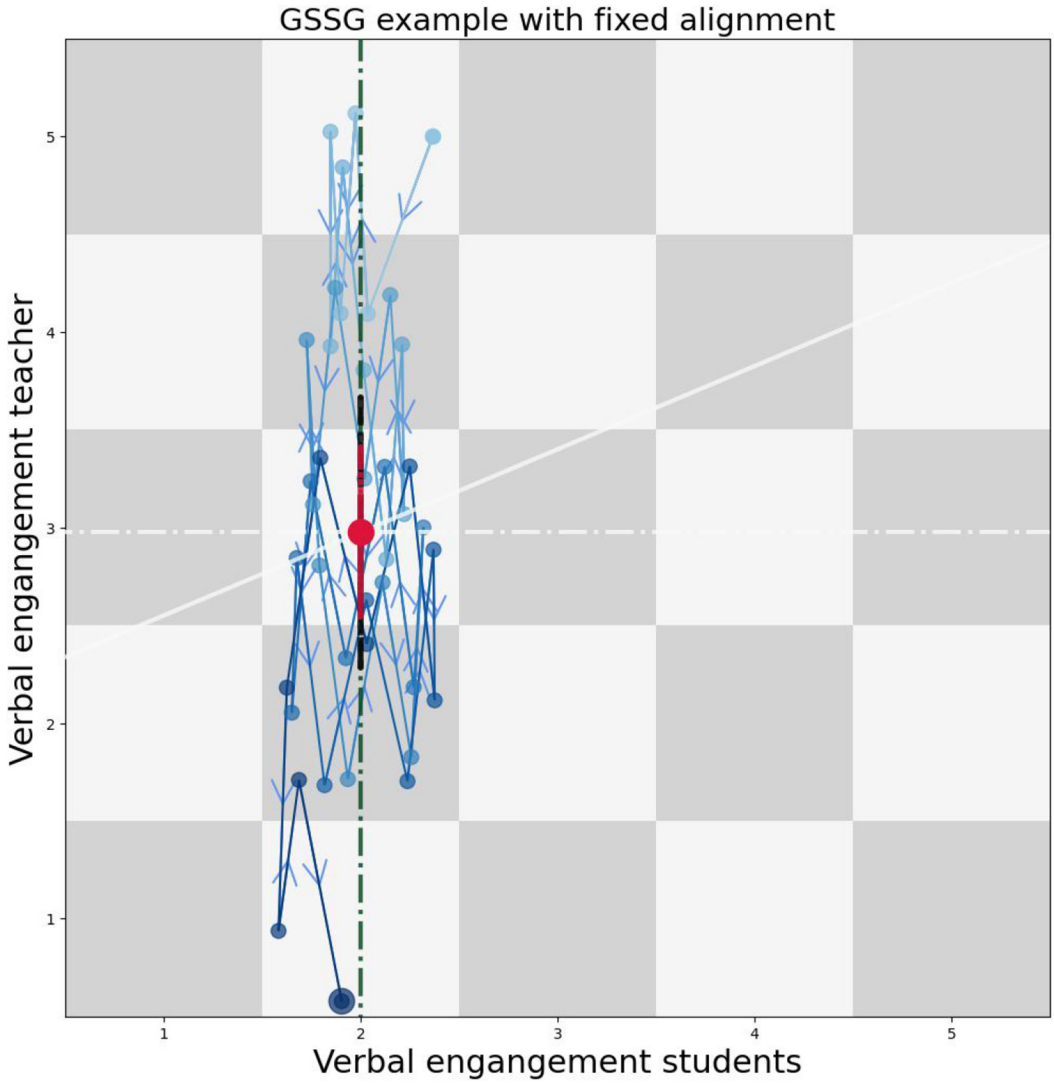

FIGURE 7
Example of a GSSG with $R^2_{vertical} = 0.85$, $R^2_{horizontal} = R^2_{linear} = 0$, $MTT = -0.5$, and $TTD = 1$.

only give insight into whether the resulting scores of the ratings are usable (Sun, 2011) and not why they might not be usable. Time-dependent ratings, like in the GSSG method, are not unrelated because they are time-dependent, so traditional parameters for rater disagreement are less appropriate. Furthermore, since we have time-dependent ratings and thus can calculate parameters for rater disagreement for each time instance, we see the potential for further insights into the rater disagreement and the influence of this rater disagreement on global assessments. In this chapter, we will propose an extension to reflect the influence of rater disagreement in our parameters, which the State Space Grid method does not. Furthermore, we propose a way to determine why a rater disagreement is high and whether there are specific events during the rated time with a low rater disagreement or if all ratings have an equal rater disagreement.

First, we look at the rater disagreement's reflection in our parameters. The idea is that ratings with a low rater disagreement contribute more than ratings with a high rater disagreement because a rating with a high disagreement is less reliable. We include this contribution of the rater disagreement through specific weights in our summations for each time instance. Thus, we need to calculate the rater disagreement for each time instance. We calculate the result of a rating for a specific time j through each rater's ratings of this specific time j through the mean value over all raters. Thus, the error $\Delta x_{i,j}$ for indicator i of the rating in time instance j is the error of the mean value over all raters' k error:

$$\Delta x_{i,j} = \frac{1}{\sqrt{p}} \sqrt{\frac{1}{p} \sum_{k=1}^{p} \left(x_{i,j}^{k} - x_{i,j}\right)^2}. \tag{11}$$

where p is the total number of raters.

Now, we can use this error of each time step as a weight for the calculations of our parameters. Since this extension with weights is similar for all parameters, we will give two examples. For further details, see: https://osf.io/eczxn/?view_only=2abaed19e3ed4d40a0ca4a8001a32512.

The first type of parameter is a parameter that depends on one rating $x_{i,j}$ for each step in the sum. We will use the Mean Standard

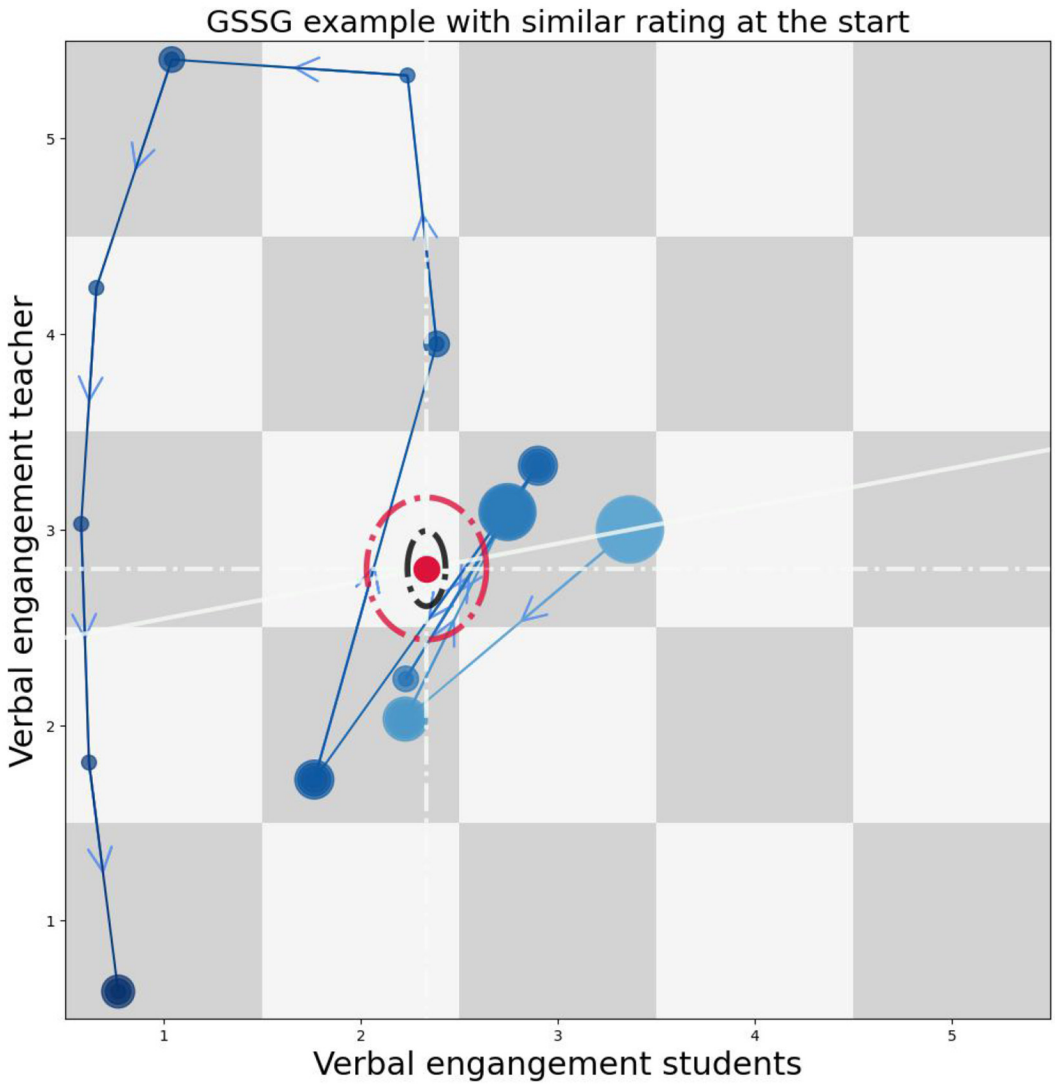

FIGURE 8
Example of a GSSG with beginning similarity $BS = 0.81$, ending similarity $ES = 0.55$, Beginning similarity Trend $BST = -0.04$, and ending similarity Trend $EST = -0.20$.

Deviation *MSD* as an example here. We adjust the location of the global attractor, the Standard Deviation Difference *SDD*, the Beginning Similarity *BS*, and the Ending Similarity *ES* similarly. We weight each part of the sum with the error $\Delta x_{(i,j)}$ and normalize the sum through a summation of reciprocate all weights $\Delta x_{(i,j)}$. Thus, the new version of the Mean Standard Deviation *MSD* for multiple raters is:

$$MSD = \frac{1}{d}\sum_{i=1}^{d}\frac{2}{o_i}\sqrt{\frac{1}{\sum_{i=1}^{n}\frac{1}{\Delta x_{i,j}}}\sum_{j=1}^{n}\left(\frac{\overline{x_i}-x_{i,j}}{\Delta x_{i,j}}\right)^2}. \qquad (12)$$

There can be cases where all raters have the same rating, resulting in $\Delta x_{i,j} = 0$. In this approach, for a weighed parameter, such a case for $\Delta x_{i,j} = 0$ would lead to a division by zero. We include this case where the error is zero as by setting $\Delta x_{i,j} = \overline{\Delta x_i}$. Through the normalization in front of the sum those cases contribute more to the sum than all ratings with an error. Through this approach, we can weight a rating with rater disagreement less than a rating without rater disagreement.

The second parameter type includes calculation, which depends on two successive ratings ($x_{i,j}$ and $x_{i,j+1}$) in each sum part. These types of parameters include the Mean Travel Distance *MTD*, the Travel Distance Difference *TDD*, the Mean Travel Tendency *MTT*, the Travel Tendency Difference *TTD*, the Beginning Similarity Trend *BST*, and the Ending Similarity Trend *EST*. Similar to the parameter types, which depend only on one rating in the sum, we divide by each error of the dependent rating and normalize the sum with the mean of all errors. We also set $\Delta x_{i,j} = \overline{\Delta x_i}$ if there is no rater disagreement, which lets us weigh these ratings more than those with an error. We give an example of this adjustment for the Mean Travel Tendency *MTT* because we adjust all other parameters similarly. Thus, to include the rater disagreement

in the Mean Travel Tendency MTT, we adjust its calculation as follows:

$$MTD = \frac{1}{\sqrt{d}}\Big(\sum_{i=1}^{d}\sum_{j=1}^{n-1}\frac{1}{\Delta x_{i,j}+\Delta x_{i,j+1}}\Big)^{-1}\sum_{j=1}^{n-1}\frac{\left|(x_{1,j},\ldots x_{d,j})-(x_{1,j+1},\ldots,x_{d,j+1})\right|}{\sum_{i=1}^{d}\Delta x_{i,j}+\Delta x_{i,j+1}}. \tag{13}$$

Finally, we have to adjust the parameters for the goodness of fit. Since a R^2 test only considers the position of each rating and not their rater disagreement, we use a more fitting test instead. A typical test for this case is the χ^2 test, which considers the difference between the error and the distance between the rating and the fit (Berendsen, 2011). Similar to the previous adjustments, this test divides the sum parts by the error of each rating in the following way:

$$\chi^2 = \frac{1}{n}\sum_{j=1}^{n}\frac{(\text{distance between fit and rating } j)^2}{\text{error of rating } j}. \tag{14}$$

Such a calculation leads to the same problem: dividing by an error of zero if the rating has no rater disagreement. This problem is why we normalize the sum with the mean of the error and set the error of the rating j to the mean value of this rating. This approach leads to the following adjustment for the χ^2 test:

$$\chi^2 = \frac{\overline{\text{error of all ratings}}}{n}\sum_{j=1}^{n}\frac{(\text{distance between fit and rating } j)^2}{\text{error of rating } j}. \tag{15}$$

Now, we can include the rater disagreement directly in our parameters, which State Space Grids does not include. In addition to the adjusted calculation, we want to represent the rater's disagreement visually. To show the rater disagreement for each time step, we use error bars which show the rater disagreement of each time step as calculated in Equation 11 (Barde and Barde, 2012). Adjusting our parameters and ratings to reflect the rater disagreement leads to GSSG plots like Figure 9. In Figure 9, we can see at first glance that the rating of both indicators has a low rater disagreement because there are only four error bars. Furthermore, the only rater disagreement happens in the time step around the point when students and the teacher both speak more than 60 % of the time. Thus, the raters disagree on how they rate these three steps in this situation. This rater disagreement also leads to a slight change in the ellipses. In comparison to Figure 4, the black ellipse is more spread out on the y-axis because the downward jump from $(1, 3)$ to $(1, 1)$ as well as the upward jump back to $(1, 3)$ do in more than the sideward jumps on the right side. The same applies to the influence of the $(1, 1)$ rating on the red ellipse, which is why the red ellipse is also more spread out on the y-axis.

The approach with error bars for the rater disagreement not only gives us an insight that the rater disagreement is low ($\overline{\Delta x_1} = \overline{\Delta x_2} = 0.01$) but also shows us, that there is only one specific situation where they disagree. We can also prove this disagreement with calculations by calculating the mean error over all errors. In this case, the error of the mean over all errors is for both indicators 0.03 and higher than the mean over all errors (0.01) itself. Thus, there have to be only a few situations with a high rater disagreement, which leads to the variability for the overall rater disagreement. Such a case tells us that the data is usable since the rater disagreement is low, but we have to be careful if we want to argue that the teacher and the students speak more in the first half of the time.

Ratings, where there is no such specific moment of rater disagreement, have equally sized error bars and thus a higher mean value of all errors than an error of this mean. Such a case tells us that if the mean over all errors ($\overline{\Delta x_i}$) is still low, the data has sufficiently low rater disagreement, and there is no specific time frame where we have to be careful in arguing about the speech distribution.

In contrast to the State Space Grid method, we can now directly include the rater disagreement's influence in our global assessments in two ways. On the one hand, we reflect the rater disagreement in our global assessment for dynamic patterns and variability. On the other hand, we have visualization through error bars that show us if the rater disagreement is sufficiently low and when and to what extent the raters disagree.

4 The analysis of transition changes with GSSG

This chapter will show how we can use the GSSG method to analyze transition changes between two lesson states. Looking at the transition change in Figure 2, we can now implement all new parameters and visualizations of the GSSG method to the two states leading to a GSSG plot like Figure 10.

We give each state a different color scheme to visualize the different states. The first state, where the teacher speaks most of the time, has blue ratings. The second state, where most of the time the students speak, has a violet color scheme. Finally, the time step between the first and the last state is orange.

At first glance, we can see that the global attractor in the first state, where the teacher speaks most of the time, has more variability in its ratings of the speech distribution than in the second state, where the students speak most of the time because the ellipses are bigger in the first state. Furthermore, the rater disagreement is higher in the first stage because the global attractor and the ratings have bigger error bars. This rater disagreement is similar in all time steps and is not due to a specific time interval because all error bars have mostly equal sizes in the first state. However, the raters disagree less when the teacher speaks less, which indicates that the disagreement of the raters happens due to the teacher speaking almost all the time. Additionally, we can see that whenever the teacher speaks more, the students speak less in the first stage (linear fit is green), and this is not the case in the second state (linear fit is white).

Furthermore, using the parameters of the GSSG method, we can find the following changes through the transition between the states:

1. The speech distribution changes more frequently when the teacher speaks mostly than later when the students speak most of the time ($MTD = 0.539 \rightarrow 0.284$). In both states, the teacher changes his behavior more frequently than the students ($TDD = 0.013 \rightarrow 0.03$).

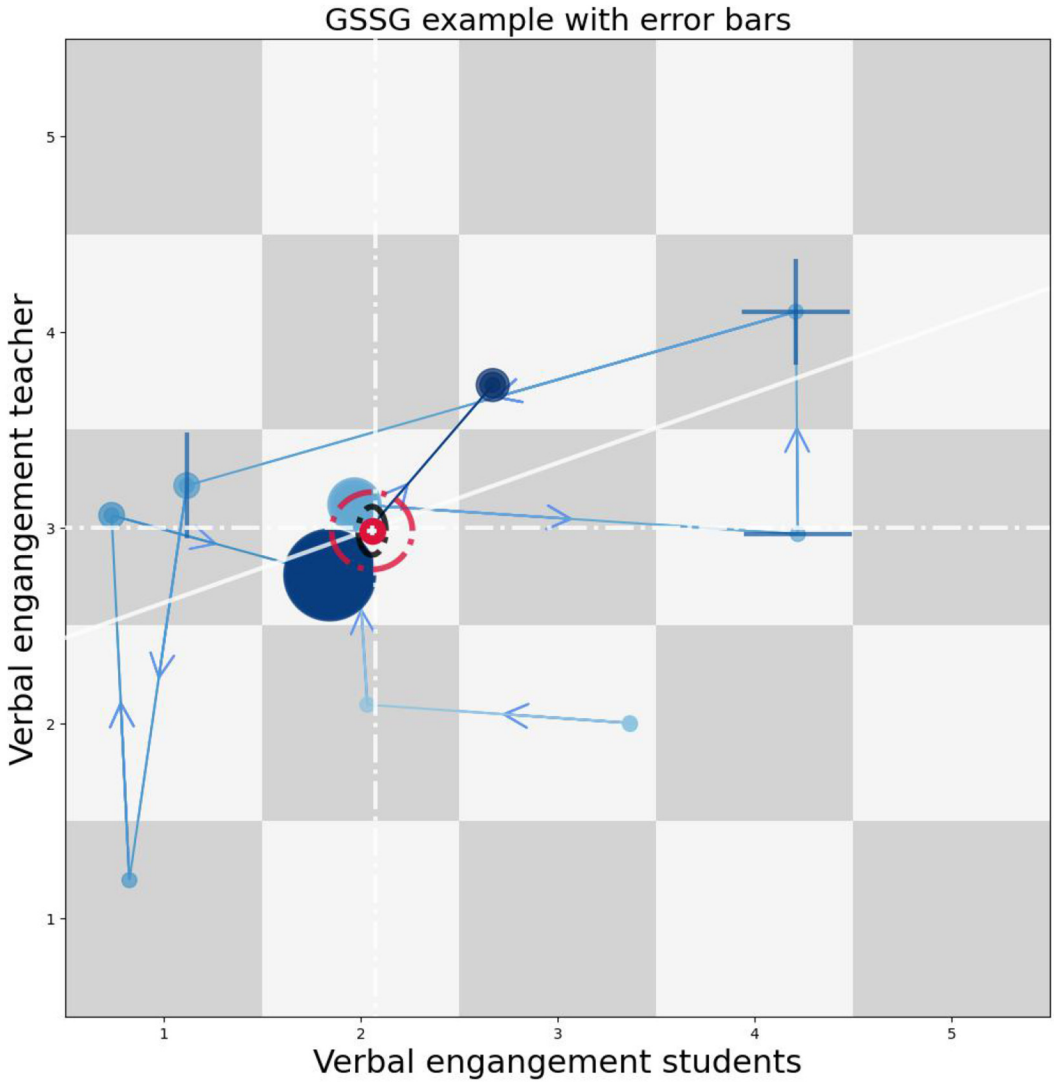

FIGURE 9
Example of a GSSG with error bars and $\overline{\Delta x_1} = \overline{\Delta x_2} = 0.01 \pm 0.03$.

2. Overall, the speech distribution is similar to the global attractor, but we need to consider the changes in the speech distribution ($MSD = 0.306 \rightarrow 0.246$). However, the students and the teacher have a similar influence on this overall variability ($SDD \approx 0$ in both states).
3. The amount of spoken time decreases in the first stage and increases in the second state ($MTT = -0.19 \rightarrow 0.089$). These two changes mostly happen because the teacher tends to speak less in the first state as time passes and more in the second ($TTD = 0.255 \rightarrow 0.178$).
4. The speech distributions are very different in both states ($BS \approx ES \approx 0$) because the teacher speaks more than the students in the first state and the students speak more than the teacher in the second state ($BST \approx EST \approx -0.4 \rightarrow 0.4$).
5. Thus, the global attractors can represent the two states, but we must consider the frequent changes, decreases, and increases in the speech distribution.

Looking at this example, the GSSG method can give further and deeper insights into transition changes between different states than the SSG method. In contrast to the SSG, which specializes in changes between two specific speech distributions, we can give further insights into the overall dynamic change of the speech distribution and the influence of this dynamic on the speech distribution of the whole lesson.

5 Practical applicability and constraints

In this chapter, we will provide a short guide for the practical application of the GSSG method through an example with real-world data of a physics lesson in Figure 11. Through this example, we will present the practical applicability and discuss constraints and limitations that researchers have to consider when using this new method.

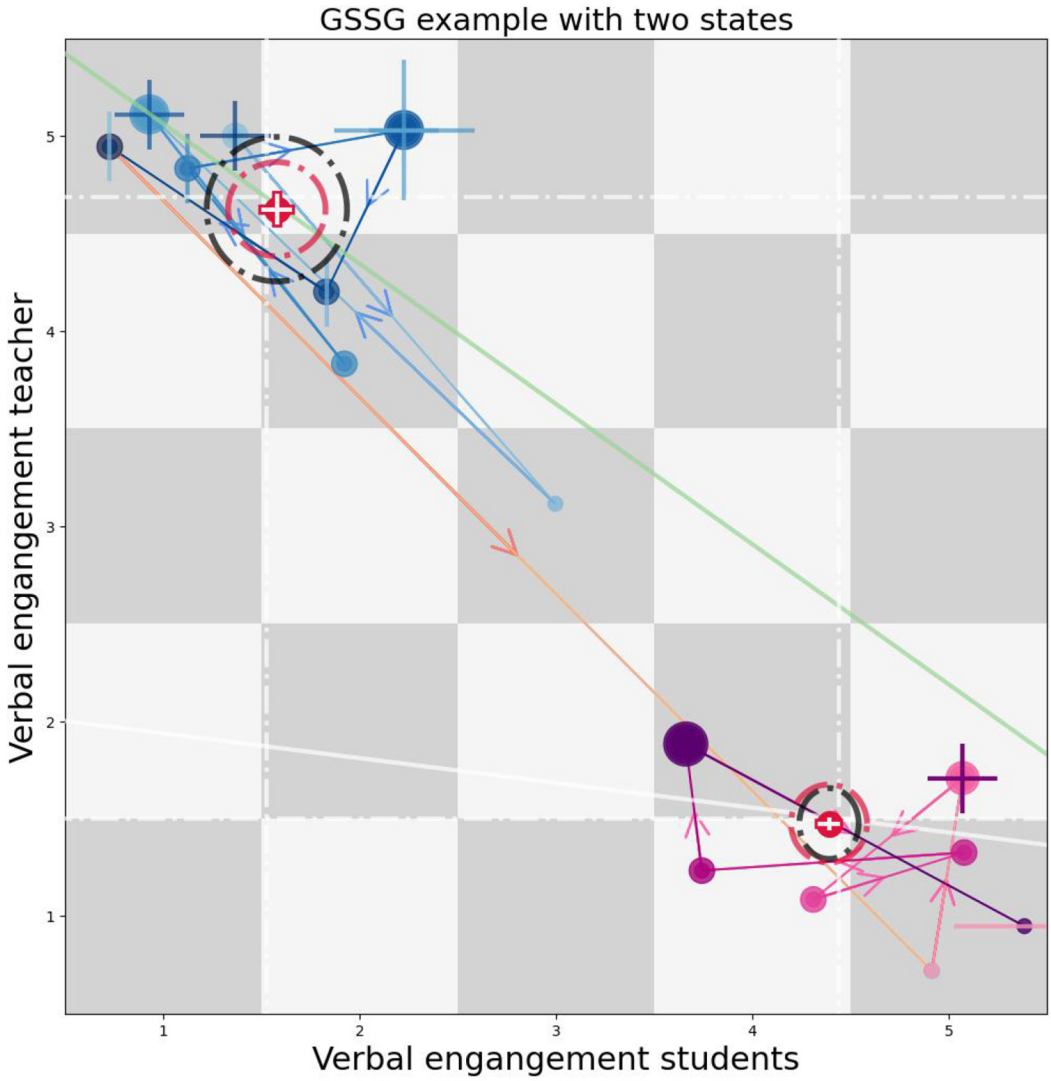

FIGURE 10
Example of a GSSG with two states.

In this second example, we analyze a physics lesson where the students measured the acceleration due to gravity with a free-fall experiment in groups. During the 45-min-long lesson, the teacher was asked to intervene in classroom disruptions by going to the groups where the disruption happened and addressing the issue there. In this example, this kind of intervention happened once from min 24 to min 31. During this time, the teacher went to three groups and intervened in the disruptions in those groups. Our goal is to provide an easy example by measuring the impact of the teacher's intervention on the class. Since this is a simple example to illustrate the practical applicability as well as the constraints and limitations, we will only analyze two indicators for the class. For a more precise understanding of how teacher interventions influence class dynamics, future studies should employ a multidimensional approach, integrating various indicators to capture the complexity of the class reaction to the intervention. We will give a few examples of such additional indicators later.

For this example, we will look at these two indicators: the first indicator measures the number of students participating in the class activity, and the second indicator measures the number of disruption types occurring during each minute. We rate the first indicator for participation similarly to the indicators of the speech distribution. We give a "1" if 20 percent or less of the students participate in the assigned class activity, a "2" if 21–40 percent of the students participate, a "3" for 41%–60%, a "4" for 61%–80%, and a "5" for 81 percent or higher.

The indicator for the disruption types has a different approach. For this indicator, we count the number of disruption types occurring during each minute. We differentiate four types of disruptions: students talking about non-related topics, interruptions by the teacher, verbal conflicts, and non-verbal conflicts. During the rating process, these four types of disruptions proved to describe all types of disruptions that occurred. Thus, additional disruption types were not necessary. For easy comparison between the two indicators, we formulate the indicator so that high values indicate "good" behavior. Thus, we give a "5" if no disruption occurs in the rated minute, a "4" if only one disruption type occurs in the rated minute, a "3" if two different

disruption types occur in the rated minute, a "2" for three types, and a "1" if all four types occur.

Since we know when the teacher intervenes, we can separate the lesson into three sections by illustrating in the GSSG plot where each time section begins: a first section where not all students are participating and a few disruptions occur (illustrated in blue in the bottom half of the plot), an intermediate section where more disruption types occur and the teacher intervenes [illustrated in red going from (1,3) to (1,2) and then to (3,3)], and a second section where the students participate more and fewer disruption types occur (illustrated in violet in the top half of the plot). Our goal in this example is to measure the impact of the teacher's intervention on the number of disruption types and the number of participating students using the GSSG method. This is why we will compare the first time section (blue) with the second time section (violet).

Following this rating scheme, twenty trained raters rated the first indicator, and five highly trained raters rated the second indicator. The different number of raters is due to the project design, but it is not important for this example. The focus is on analyzing rater agreement. It is possible to use these two indicators with only one or two raters. The results are presented in Figure 11, which includes a GSSG plot with all parameters. Such a complex figure can be overwhelming at first sight, but we will go through the analysis step by step.

First, before we take a look at the ratings themselves, we will look at the possible constraint of rater disagreement, which is in the bottom right of the figure. Looking at the errors of the indicator displayed at the x-axis, we find $\Delta x = 0.06 \pm 0.05 \rightarrow 0.04 \pm 0.04$. These four values tell us that the rater disagreement is low for the time sections, and if the raters disagree, they disagree only slightly at a few minutes of the lesson's duration, because the error of the average error is as high as the average error itself. Thus, we can use the GSSG method for this indicator without considering further effects of the disagreement. However, this is different for the indicator displayed at the y-axis. Here, the disagreement is higher with $\Delta y = 0.27 \pm 0.01 \rightarrow 0.19 \pm 0.01$, which tells us that there is a general but low disagreement for the whole duration of the lesson. We can still use these ratings since they are well below 0.5, where the disagreement makes it questionable to differentiate between each step of the rating scheme. However, if we want to compare an overall change in participation, we need to consider this amount of disagreement for the following analysis. For example, if the difference in the overall change in participation is 0.15 and, thus, lower than the disagreement, such a change of 0.15 can not be considered significant.

Now we can take a first look at the time development of the ratings. Looking at the sections where each color of the rating, we find that the violet ratings are above and more on the right side than the blue ratings, because the teacher's intervention (red ratings) affected the participation and the types of disruptions that occurred. These two sections of the plot overlap slightly at the end of the lesson, and we will consider this overlap later. To explore the impact of the teacher's intervention on the class, we will examine the changes in indicators between the blue section and the violet section.

The first comparison we can make is through the number of ratings in each section. In the top right of Figure 11, we can find that 14 ratings are within the first section (blue) and 24 are in the third section (violet). Since we included the whole lesson of 45 min, the duration of the intervention in the red ratings consists of $45 - (14 + 24) = 7$ ratings. Since each rating has a duration of 1 min, we know that the teacher's intervention took 7 min, where the teacher went from group to group working on the experiments to intervene in the disruptions. Furthermore, the positive effect of the intervention lasted 22 min, as participation decreased in the last 2 min. This exclusion of the last 2 min has some further implications, which we will look into when we consider the limitations of the GSSG method.

Next, we interpret the variability parameters. In the first two lines of the variability parameters, we can see the average change in the ratings before and after the intervention. Thus, the participation increases on average from 2.66 ± 0.21 to 3.99 ± 0.11, representing an increase of 1.33 or an increase by 50%. Furthermore, on average 0.23 fewer disruption types occurred. These two changes are higher than the sum of their measurement errors (0.31 and 0.06) and also higher than the average error of each rating ($0.27 \rightarrow 0.19$ and $0.06 \rightarrow 0.04$). Hence, the disagreement of the raters does not substantially affect the measure of change in participation. However, we still do not know if the two mean values we compared can represent the lesson's section under study. We can answer this question by looking at the MSD, which is low enough ($MSD_1 = 0.432$, $MSD_2 = 0.444$) for the mean values to represent the overall sections of the lesson. Thus, we can compare these two sections with the mean value. A general rule of thumb for the MSD is that its value must be less than 0.5 for a rating system with ratings from 1 to 5. However, the GSSG method still lacks concrete reference values, which is a limitation in the current state of the GSSG method. Thus, we can compare these two sections with the mean value.

Based on the variability parameter, we found the following results for the impact of the teacher's intervention on the disruptions in this particular lesson: The participation increases significantly by 50%, and the disruption types decrease significantly by 0.23 disruption types on average. Thus, we can not only analyze if the intervention was successful, but also to what extent (50% better participation and 0.23 less disruption types).

There are two more insights we can gain through the variability parameters. The first one is that the average change in ratings from minute-to-minute decreases ($MTD = 1.132 \rightarrow 0.929$). Hence, the teacher's intervention not only increases the participation and decreases the number of disruption types, but also increases the minute-to-minute stability of the participation and the disruption types by 18%. However, we do not know if this change in stability only happens because one indicator has an increased stability or if both indicators gain more stability. This question can be answered with the SDD and TDD values. We find positive and barely changed values for both time sections, which tells us that one of the two indicators is more variable than the other. Looking at the ellipses, we can see that the disruption types are more variable. However, since the orientation of the ellipse remains the same for both time sections as well as the SDD and TDD values, the change in MTD is meaningful for both indicators and is not caused by just one indicator.

In the next step of the analysis through the GSSG method, we look at additional effects of the intervention through a change in

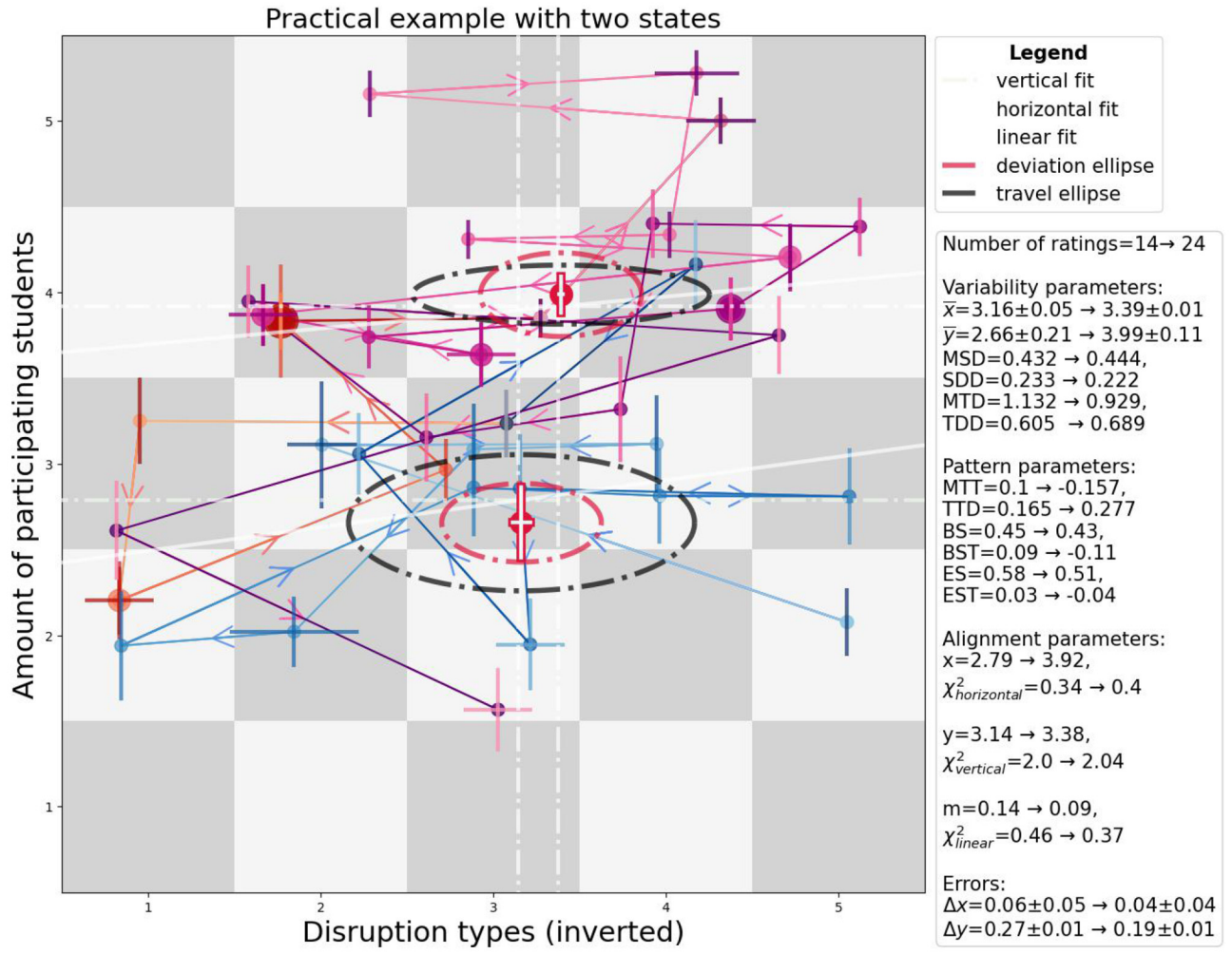

FIGURE 11
Example of real data with transition changes and all parameters.

rating pattern or alignments. Looking at the pattern parameters, we find nothing of interest. The MTT, TTD, BS, BST, ES, and EST remain low. For the MTT, BST, and EST, a general rule of thumb is that the absolute value of the MTT, BST, and EST needs to be above 0.20. For the BS and ES values, we consider a value higher than 0.8 to be high enough. However, these recommendations still need verification, which is part of the future development of the method. Thus, the teacher's intervention did not cause a minute-to-minute increase or decrease in the ratings or any kind of similarity. In this case, it is a sign that the separation of the time sections is meaningful, because there is no additional trend in the rating which could suggest the need for more detailed time sections. For example, a high MTT for both time sections would indicate that the participation steadily increases or decreases, which would make no sense if we want to measure a sudden increase in the participation due to the teacher's intervention.

The alignment parameters show similar results. Looking at the χ^2 values, we find no significant goodness of fit, which means we can also rule out any additional direct linear increase in ratings (χ^2_{linear} high enough). Furthermore, we can also rule out a case where, for example, the participation does not change but the number of disruption types does ($\chi^2_{horizontal}$ and $\chi^2_{vertical}$ high enough).

In summary, our results on the question to what extend teachers' intervention affects the disruptions in class show: There is a reliable (low enough MSD and rater disagreement) increase by 50% (ratio between the mean values) for the participation of the students and on average 0.23 fewer disruption types (difference in mean value) after the teacher's interventions. Additionally, after the interventions, the minute-to-minute stability of the class workflow increased by 18% (ratio between the MTD values), regarding the participation and the disruption types. Furthermore, we can also rule out any additional effects in the ratings pattern (MTT, TTD, BS, ES, BST, EST low enough) or alignment (χ^2 values high enough) that might affect this finding.

Apart from the specific constraints of the GSSG method's applicability for this example, there are also general constraints and limitations. As we have seen in Figure 11, it is important which minutes we include in each time section. For example, if we only want to look at the duration where the teacher's intervention is

effective, we have to exclude the last 2 min. In this case, shorter time sections could be more meaningful if we only want to compare the time when class participation remains high to the time section before the intervention. Furthermore, the duration of each rating also sets a limitation on the effects we can find. For example, one rating for every 5 min would make the comparison before and after an intervention less meaningful, especially if the intervention only lasts 1 min.

While shorter times for each rating decreases this limitation, rating multiple indicators many times with a short duration is demanding for raters and can increase rater disagreement. However, it is also possible to use the GSSG method with one or two trained raters.

Since the GSSG method relies on many calculations of distances in each dimension, the definition of the indicators is more limited than for traditional methods. It is essential that the difference between a "1" and a "2" is equidistant to the difference between a "4" and a "5" if we want to measure any kind of average over the ratings. Thus, indicators that count a quantity (like time, number, or proportion of students) are preferable. Likert-scaled indicators are also usable, but the categories of the rating scheme should be similar to metric measurement levels. However, since the GSSG method can include more than just two dimensions, adding one Likert-scaled indicator to a set of multiple other indicators is possible. This way, it is also possible to measure impacts on a more general class workflow through multiple aspects, like participating students, students speaking about content-related subjects, students paying attention, or students looking at subject-related content (like books).

Another constraint of the GSSG method is that determining whether a value is sufficiently low or high to be considered relevant, for example, whether $MTT = 0.1$ is sufficiently low, is challenging and requires experience with the method. Reference values could be a way to overcome this constraint, which we are currently working on.

Analyzing a single lesson with the GSSG method can provide a deep insight into the lesson's dynamics. However, a meaningful analysis of the impact of a teacher's intervention on student participation and the types of disruptions that occur needs multiple lessons with varying intervention techniques. This way, the GSSG method could compare the impact of different intervention types on the overall change and stability of class participation.

6 Conclusion and outlook

We proposed a comprehensive extension of the SSG method to give a potential solution to the demand to expand existing methods (Dirk and Nett, 2022; Dietrich et al., 2022; Pekrun and Marsh, 2022; Moeller et al., 2022) to analyze dynamics of teaching and furthermore to consider time-dependent variations within a global assessment of a whole lesson, that are typically not considered (Praetorius et al., 2020).

We showed how newly developed numerical parameters could give further insights into the dynamic interactions of indicators. We illustrated these insights with an example of two indicators which we rated each minute for the entire duration of a lesson. However, we can also use event sampling to rate and use the indicators in this method by weighing each event accordingly. Furthermore, we can also use the GSSG method with more than just two indicators because our parameters give all insights without the need for a graphical representation. So, there is no limit to the number of indicators we can assess at the same time.

The GSSG method enables the measurement of the frequency and manner in which indicators change over time, which is impossible with a single global rating. We can determine how much influence trend effects have (e.g., if the indicator ratings increase or become more similar over time) and how big the difference in each change is. Furthermore, we can simultaneously compare these changes to an unlimited number of other indicators and compare these changes between two different states.

Using the GSSG method, future research can, for example, find out how much and in what way a teacher can increase the amount students speak. This method gives a new perspective to teaching research since we not only look at linear relations. For example, in cases where the teacher speaks as much as the students could prove to lead to a speech distribution where the students speak more and more over time. So, a teacher who wants to have the students speak more could specifically target short situations where he and the students speak for an equal amount of time by, for example, waiting longer for responses.

The GSSG method could also provide more insight into the field of interventions for classroom disruptions. Comparing different aspects of student workflow before and after an intervention of classroom disruptions can gain valuable insights into the increase in overall student workflow and workflow stability. This way, comparing the impact of different intervention techniques can provide empirically based tips on when to use which intervention technique.

Additional fields for a similar approach with the GSSG method could be the impact and effectiveness on the students' workflow through (a) teacher explanation styles, like frequent and appreciative student-teacher interactions, (b) teacher agency and communion, (c) collaborative behaviors, like students effectively working as a team, and (d) different kinds of lessons, like comparing experimental lessons and theory lessons, or comparing natural science lessons with language lessons.

We also take time-dependent rater disagreement into account and gain further insight into the effects of rater disagreement, which would not be possible with global ratings. For example, the GSSG method can determine if raters have a higher disagreement in the first 2 min of a lesson than in the rest, which gives potentially more insight into the reasons for rater disagreement. This way, rater training can be improved and made even more specific in order to reduce rater bias.

To make the GSSG method accessible to all, we are making the Python code freely available: https://osf.io/eczxn/?view_only=1baf125d5bc44e2fa065a8bd6022621e. This code generates two-dimensional GSSG plots and calculates all shown parameters. Additionally, it can separate between two states and calculate the results of the parameters for each state.

While the GSSG method can provide valuable insights into the dynamics of lessons, it should not be regarded as a replacement for traditional methods, like questionnaires or interviews. Instead, using a mixed method approach by combining the analysis of dynamics in the classroom with the GSSG method and interviews and/or questionnaires of the teachers and students, for example, can gain further and deeper insights into the reasoning behind dynamic decisions during the lesson. For the advancement of research, recognizing the capacity of each method can lead to mutual enhancement and further potential for researchers rather than exclusion.

At the current state of the development of the GSSG method, it remains challenging to interpret a given GSSG plot through its parameters because of the lack of meaningful reference values. For example, interpreting whether 0.607 for the result of the Mean Travel Distance is considered high or low is very challenging and time-consuming for beginners using this method. Thus, in the next step, we will develop reference values for GSSG plots and provide a detailed guide for interpreting the results of each parameter. This way, it will be possible to further categorize types of GSSG plots quickly, only through the GSSG parameters.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/eczxn/?view_only=2abaed19e3ed4d40a0ca4a8001a32512.

Author contributions

NL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AP: Investigation, Visualization, Writing – original draft, Writing – review & editing. SW: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 233630050 – TRR 146.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Barde, M., and Barde, P. (2012). What to use to express the variability of data: standard deviation or standard error of mean? *Perspect. Clin. Res.* 3, 113–116. doi: 10.4103/2229-3485.100662

Bell, C., Dobbelaer, M., Klette, K., and Visscher, A. (2019). Qualities of classroom observation systems. *Sch. Eff. Sch. Improv.* 30, 3–29. doi: 10.1080/09243453.2018.1539014

Belova, N., and Eilks, I. (2015). Learning with and about advertising in chemistry education with a lesson plan on natural cosmetics - a case study. *Chem. Educ. Res. Pract.* 16, 578–588. doi: 10.1039/C5RP00035A

Bento, T., Ribeiro, A., Salgado, J., Mendes, I., and Gonçalves, M. (2014). The narrative model of therapeutic change: an exploratory study tracking innovative moments and protonarratives using state space grids. *J. Constr. Psychol.* 27, 41–58. doi: 10.1080/10720537.2014.850373

Berendsen, H. J. (2011). *A Student's Guide to Data and Error Analysis*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511921247

Blank, R. (1993). Developing a system of education indicators: selecting, implementing, and reporting indicators. *Educ. Eval. Policy Anal.* 15, 65–80. doi: 10.3102/01623737015001065

Brennan, R. L., and Prediger, D. J. (1981). Coefficient kappa: some uses, misuses, and alternatives. *Educ. Psychol. Meas.* 41, 687–699. doi: 10.1177/001316448104100307

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., et al. (2019a). "Measuring teacher effectiveness across time: what does TIMSS reveal about education system level trends?" in *Teaching for Excellence and Equity: Analyzing Teacher Characteristics, Behaviors and Student Outcomes with TIMSS*, eds. N. Burroughs, J. Gardner, Y. Lee, S. Guo, I. Touitou, K. Jansen, et al.(Cham: Springer International Publishing), 29–45. doi: 10.1007/978-3-030-16151-4_4

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., et al. (2019b). *Teaching for Excellence and Equity: Analyzing Teacher Characteristics, Behaviors and Student Outcomes with TIMSS*. New York, NY: Springer Nature. doi: 10.1007/978-3-030-16151-4

Cerbin, W., and Kopp, B. (2006). Lesson study as a model for building pedagogical knowledge and improving teaching. *Int. J. Teach. Learn. High. Educ.* 18, 250–257.

Charalambous, C., Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively. *ZDM* 50, 355–366. doi: 10.1007/s11858-018-0914-8

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 220, 37–46. doi: 10.1177/001316446002000104

Dietrich, J., Schmiedek, F., and Moeller, J. (2022). Academic motivation and emotions are experienced in learning situations, so let's study them. Introduction to the special issue. *Learn. Instr.* 81:101623. doi: 10.1016/j.learninstruc.2022.101623

Dilshad, M., and Iqbal, H. (2010). Quality indicators in teacher education programmes. *Pak. J. Soc. Sci.* 30, 401–411.

Dirk, J., and Nett, U. (2022). Uncovering the situational impact in educational settings: Studies on motivational and emotional experiences. *Learn. Instr.* 81:101661. doi: 10.1016/j.learninstruc.2022.101661

Dorfner, T., Förtsch, C., and Neuhaus, B. (2018). Effects of three basic dimensions of instructional quality on students' situational interest in sixth-grade biology instruction. *Learn. Instr.* 56, 42–53. doi: 10.1016/j.learninstruc.2018.03.001

Erickson, K. (2009). *State Space Grids: First Application of a Novel Methodology to Examine Coach-athlete Interactions in Competitive Youth Sport*. Kingston, ON.

Erickson, K., Côté, J., Hollenstein, T., and Deakin, J. (2011). Examining coach-athlete interactions using state space grids: an observational analysis in competitive youth sport. *Psychol. Sport Exerc.* 12, 645–654. doi: 10.1016/j.psychsport.2011.06.006

Espin, C., and Yell, M. (1994). Critical indicators of effective teaching for preservice teachers: relationship between teaching behaviors and ratings of effectiveness. *Teach. Educ. Spec. Educ.* 17, 154–169. doi: 10.1177/088840649401700303

Fisher, R. A. (1938). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Meas. Phys. Educ. Exerc. Sci.* 5, 13–34. doi: 10.1207/S15327841MPEE0501_2

Graf, S., and Agergaard, K. (2022). *Time' Matters - Time 'Matters' - Observing Instructional Quality and the Concept of Teaching*. Hveragerði.

Granic, I., and Lamey, A. V. (2002). Combining dynamic systems and multivariate analyses to compare the mother-child interactions of externalizing subtypes. *J. Abnorm. Child Psychol.* 30, 265–283. doi: 10.1023/A:1015106913866

Hattie, J. (2003). *Teachers Make a Difference: What is the Research Evidence*? Melbourne, VIC.

Heinze, A., and Erhard, M. (2006). How much time do students have to think about teacher questions? An investigation of the quick succession of teacher questions and student responses in the German mathematics classroom. *ZDM* 38, 388–398. doi: 10.1007/BF02652800

Hoekstra, F., Martin Ginis, K., Collins, D., Dinwoodie, M., Ma, J., Gaudet, S., et al. (2023). Applying state space grids methods to characterize counsellor-client interactions in a physical activity behavioural intervention for adults with disabilities. *Psychol. Sport Exerc.* 65:102350. doi: 10.1016/j.psychsport.2022.102350

Hollenstein, T. (2007). State space grids: analyzing dynamics across development. *Int. J. Behav. Dev.* 31, 384–396. doi: 10.1177/0165025407077765

Hollenstein, T. (2012). State *Space Grids. Depicting Dynamics Across Development*. New York, NY: Springer.

Hollenstein, T. (2013). *State Space Grids*. Cham: Springer. doi: 10.1007/978-1-4614-5007-8

Ito, Y. (2019). The effectiveness of a CLIL basketball lesson: a case study of Japanese junior high school CLIL. *Engl. Lang. Teach.* 12:42. doi: 10.5539/elt.v12n11p42

Lewis, M., Lamey, A., and Douglas, L. (1999). A new dynamic systems method for the analysis of early socioemotional development. *Dev. Sci.* 2, 457–475. doi: 10.1111/1467-7687.00090

Li, Y., Chen, X., and Kulm, G. (2009). Mathematics teachers' practices and thinking in lesson plan development: a case of teaching fraction division. *ZDM Math. Educ.* 41, 717–731. doi: 10.1007/s11858-009-0174-8

Lotz, M., Gabriel, K., and Lipowsky, F. (2013). Niedrig und hoch inferente Verfahren der Unterrichtsbeobachtung. Analysen zu deren gegenseitiger Validierung. *Z. Pädagogik* 59, 357–380. doi: 10.25656/01:11942

Mainhard, M., Brekelmans, M., and Wubbels, T. (2011). Coercive and supportive teacher behaviour: within- and across-lesson associations with the classroom social climate. *Learn. Instr.* 21, 345–354. doi: 10.1016/j.learninstruc.2010.03.003

Mainhard, M., Pennings, H., Wubbels, T., and Brekelmans, M. (2012). Mapping control and affiliation in teacher-student interaction with State Space Grids. *Teach. Teach. Educ.* 28, 1027–1037. doi: 10.1016/j.tate.2012.04.008

Martin, M., Von Davier, M., and Mullis, I. (2020). "Methods and procedures: TIMSS 2019 technical report," in *International Association for the Evaluation of Educational Achievement* (Boston, MA).

Maskus, R. (1976). *Unterricht Als Prozess. Dynamisch-integratives Strukturmodell*. Bad Heilbrunn: Klinkhardt.

Meinecke, A., Hemshorn De Sanchez, C., Lehmann-Willenbrock, N., and Buengeler, C. (2019). Using state space grids for modeling temporal team dynamics. *Front. Psychol.* 10:863. doi: 10.3389/fpsyg.2019.00863

Moeller, J., Viljaranta, J., Tolvanen, A., Kracke, B., and Dietrich, J. (2022). Introducing the DYNAMICS framework of moment-to-moment development in achievement motivation. *Learn. Instr.* 81:101653. doi: 10.1016/j.learninstruc.2022.101653

Mullis, I., Martin, M., Foy, P., Kelly, D., and Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Boston, MA.

Ozili, P. K. (2023). "The acceptable R-square in empirical modelling for social science research," in *Social Research Methodology and Publishing Results: A Guide to Non-Native English Speakers* (London: IGI global), 134–143. doi: 10.4018/978-1-6684-6859-3.ch009

Pekrun, R., and Marsh, H. (2022). Research on situated motivation and emotion: progress and open problems. *Learn. Instr.* 81:101664. doi: 10.1016/j.learninstruc.2022.101664

Pennings, H., Hollenstein, T. (2019). Teacher-student interactions and teacher interpersonal styles: a state space grid analysis. *J. Exp. Educ.* 88, 382–406. doi: 10.1080/00220973.2019.1578724

Pennings, H., and Mainhard, T. (2016). "Analyzing teacher-student interactions with state space grids," in *Complex Dynamical Systems in Education: Concepts, Methods and Applications* (Cham: Springer International Publishing), 233–271. doi: 10.1007/978-3-319-27577-2_12

Praetorius, A.-K., Klieme, E., Herbert, B., and Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of three basic dimensions. *ZDM Math. Educ.* 50, 407–426. doi: 10.1007/s11858-018-0918-4

Praetorius, A.-K., Martens, M., and Brinkmann, M. (2020). "Unterrichtsqualität aus Sicht der quantitativen und qualitativen Unterrichtsforschung," in *Handbuch Schulforschung* (Cham: Springer Fachmedien Wiesbaden), 1–20. doi: 10.1007/978-3-658-24734-8_40-1

Provenzi, L., Borgatti, R., Menozzi, G., and Montirosso, R. (2015). A dynamic system analysis of dyadic flexibility and stability across the Face-to-Face Still-Face procedure: application of the State Space Grid. *Infant Behav. Dev.* 38, 1–10. doi: 10.1016/j.infbeh.2014.10.001

Ribeiro, A., Bento, T., Salgado, J., Stiles, W., and Gonçalves, M. (2011). A dynamic look at narrative change in psychotherapy: a case study tracking innovative moments and protonarratives using state space grids. *Psychother. Res.* 21, 54–69. doi: 10.1080/10503307.2010.504241

Scherzinger, M., Roth, B., and Wettstein, A. (2020). Pädagogische Interaktionen als Grundbaustein der Lehrperson-Schüler*innen-Beziehung. Die Erfassung mit State Space Grids. *Unterrichtswissenschaft* 49, 303–324. doi: 10.1007/s42010-020-00089-1

Schleicher, A. (2019). *PISA 2018: Insights and Interpretations*. Paris: OECD Publishing.

Seidel, T. (2005). "Video analysis strategies of the IPN video study-a methodological overview," in *How to Run a Video Study. Technical Report of the IPN Video Study* (Münster: Waxmann), 70–78.

Shavelson, R., Mcdonnell, L., and Oakes, J. (1990). What are educational indicators and indicator systems? *Pract. Assess. Res. Eval.* 2:11. doi: 10.7275/rtkj-a222

Smit, N. (2016). *Using State Space Grids to Analyze the Dynamics of Teacher-Student Interactions in Foreign Language Classrooms*. New Jersey.

Smith, M. S. (1988). Educational indicators. *Phi Delta Kappan* 69, 487–491.

Stigler, J. W., and Ronald, G. (2000). Using video surveys to compare classrooms and teaching across cultures: examples and lessons from the TIMSS video studies. *Educ. Psychol.* 35, 87–100. doi: 10.1207/S15326985EP3502_3

Sun, S. (2011). Meta-analysis of Cohen's kappa. *Health Serv. Outcomes Res. Method* 11, 145–163. doi: 10.1007/s10742-011-0077-3

Turner, J., and Christensen, A. (2020). Using state space grids to analyze teacher-student interaction over time. *Educ. Psychol.* 55, 256–266. doi: 10.1080/00461520.2020.1793763

Vrikki, M., Warwick, P., Vermunt, J., Mercer, N., and Van Halem, N. (2017). teacher learning in the context of lesson study: a video-based analysis of teacher discussions. *Teach. Teach. Educ.* 61, 211–224. doi: 10.1016/j.tate.2016.10.014

Welsch, R., and Devlin, P. (2007). Developing preservice teachers' reflection: examining the use of video. *Act. Teach. Educ.* 28, 53–61. doi: 10.1080/01626620.2007.10463429