

# North Carolina Public Schools Scoring Study 1 2015-2016 Aggregated

*Friday 5<sup>th</sup> February, 2016*

Scoring study title: **Scoring Study 1 Aggregated Results**

Scoring study description: **OCT 2015-16: First scoring study**

Protocol: **NCEES Teacher Evaluation Rubric 0-4**

Study dates: **12/10/2015 - 1/31/2016**

Number of Observers included: **306**

Number of videos assigned: **1**

# Table of Contents

TARGET AND MODAL SCORES . . . . .	3
Distribution of Domain Scores . . . . .	4
TARGET AGREEMENT . . . . .	7
Agreement with Target Scores by Domain and by Video . . . . .	7
MODAL AGREEMENT . . . . .	10
Agreement with Modal Scores by Domain and by Video . . . . .	10

## TARGET AND MODAL SCORES

The table below lists the target scores for each video alongside the modal scores, or the most frequently given scores, across Observers for each of the videos in your scoring study. It is possible that more than one score could be considered the modal score if they share the highest frequency among your Observers. In this case, all modal scores are listed and separated by commas. For example, if an equal number of Observers provided a score of 2 and a score of 3 on a particular domain, and these were the most common scores, then the modal score would be listed as "2, 3". You may wish to review any discrepancies between target scores and modal scores with your group of Observers.

Domain Code and Description	Video V209	
	Mode	Target
ST1a Teacher leads in classroom	2	2
ST2a Teacher provides pos environm	2	2
ST2b Teacher embraces diversity	0	0
ST2c Treats students as individuals	2	1
ST2d Adapts teach to sp needs stud	0	1
ST3a Aligns instruct w/ NC stand	2	1
ST3b Knows content for specialty	2	2
ST3c Recognizes interconnectedness	2	1
ST3d Makes instruction relevant	1	0
ST4a Knows how learning takes place	2	1
ST4b Plans instruction appr to stud	2	2
ST4c Uses a variety of inst methods	2	2
ST4d Integrates technology	2	2
ST4e Develops crit think+prob solve	2	2
ST4f Helps stud work in teams	2	2
ST4g Communicates effectively	2	1
ST4h Uses variety of assess methods	2	1

Table 1: Distribution of Modal and Target Scores by Video

## Distribution of Domain Scores

### This section helps answer the following questions

- How much did scores vary across domains for each video?
- How representative is the modal score of all scores provided?

### DISTRIBUTION OF RATINGS BY DOMAIN

The following figures illustrate the distribution of scores for each video in your scoring study. The length of each bar indicates the percentage of participating Observers that provided the score indicated in the "Scores" column to the left of the bar. The modal score for each domain is represented by the longest bar and is **bolded** in the "Scores" column. The number to the right of each bar represents the number of Observers that provided that score. Observers who have not completed any scoring for a particular video are not included in the bar graph; counts of excluded Observers are

indicated at the bottom of each figure. The target scores for each domain are also indicated in the leftmost column.

### HOW TO INTERPRET THE RESULTS

These graphs give you a sense of the variability in scores provided by the Observers in your scoring study. For a given domain, if you see one long bar surrounded by much smaller bars or even just a single bar alone, this indicates that your group of Observers generally agreed on one score. Preferably, this score would match the target score. If the modal score does not match the target score, you may want to discuss this among your observer group and/or perhaps consider adjusting the target score. However, if instead you see several bars of similar length and no one bar stands out from the rest, this means that there was significant disagreement on how the domain should be scored. This lack of consensus could indicate that further training is needed for your Observers on that particular domain. In any case, if there is any disagreement you may wish to facilitate further discussions with the group to uncover why people provided different scores and come to some consensus on what the true score should be.

VIDEO V209

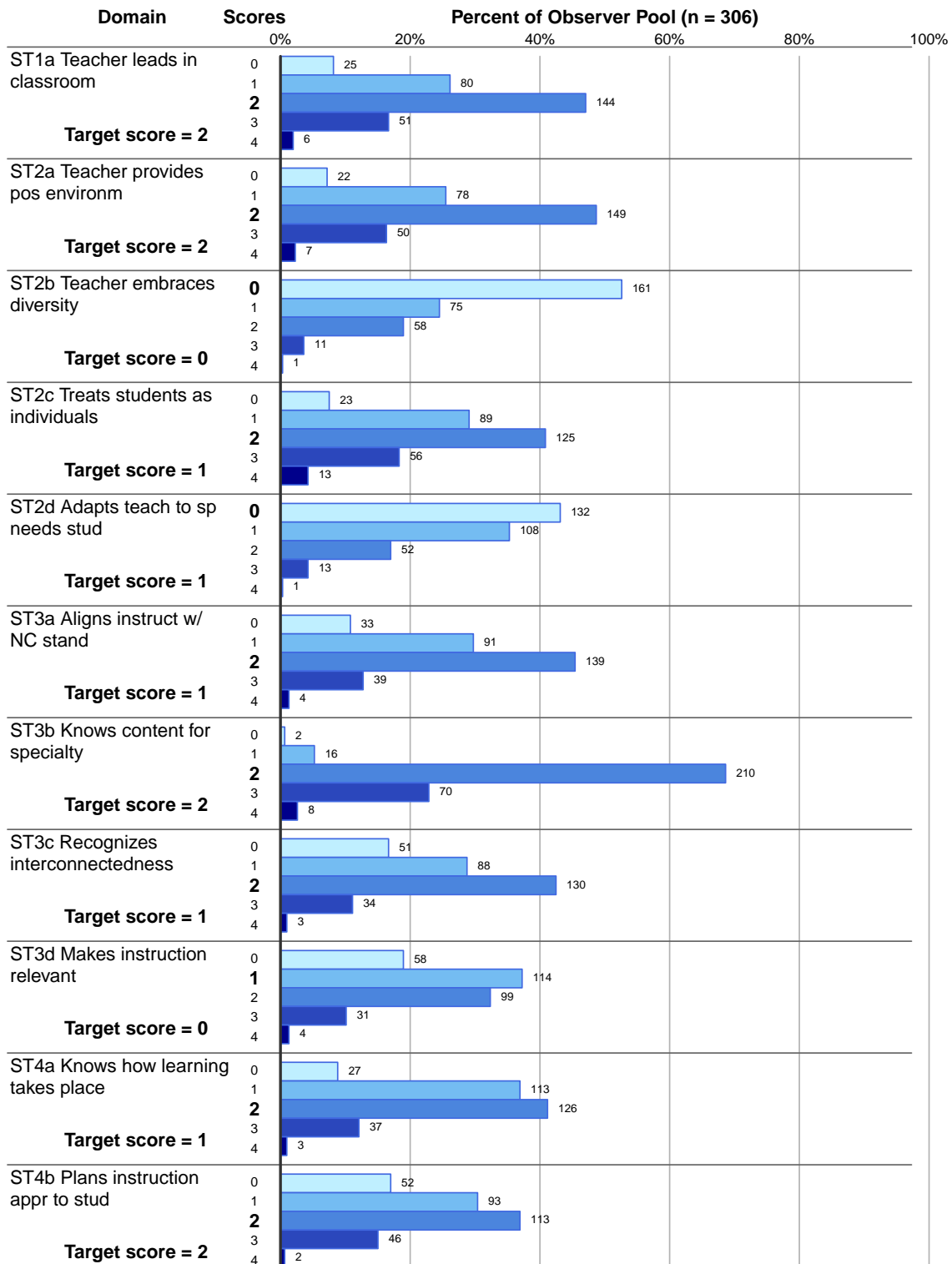


Figure 1: Distribution of Ratings by Domain for Video V209

VIDEO V209

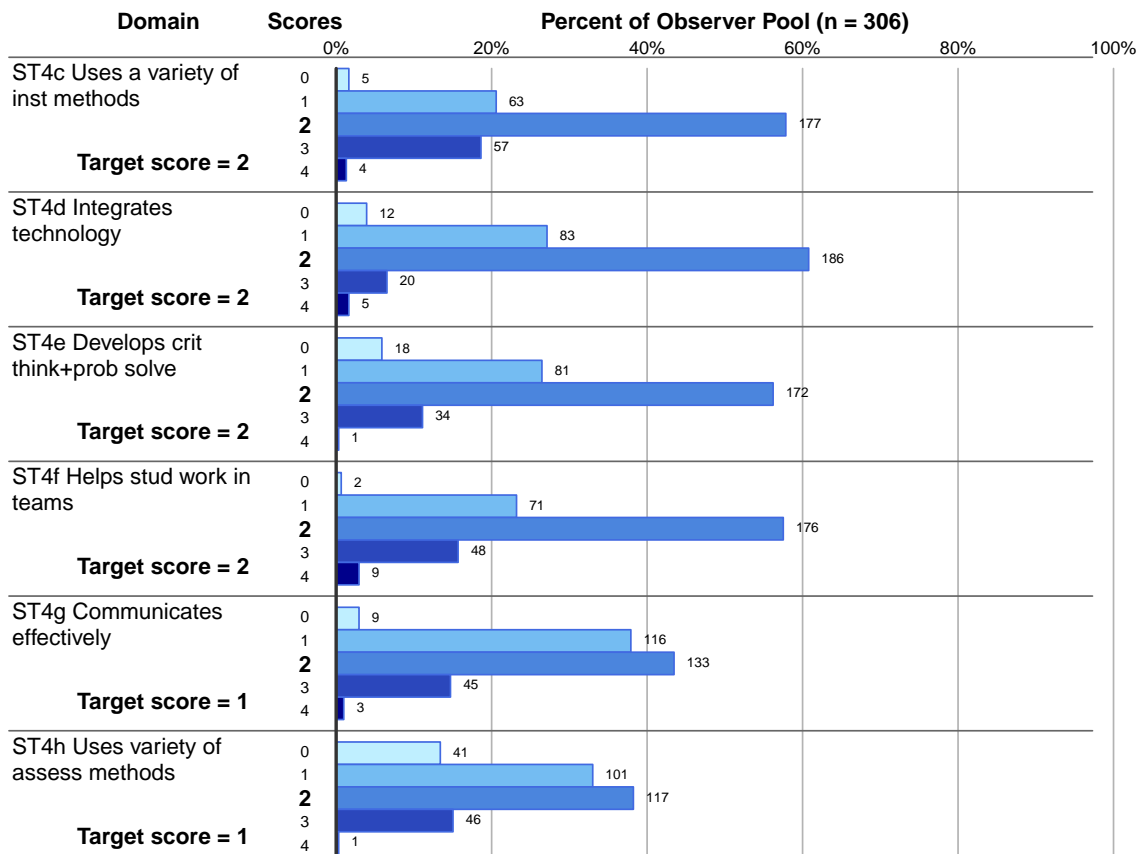


Figure 1: (Continued)

## TARGET AGREEMENT

### Agreement with Target Scores by Domain and by Video

#### This section helps answer the following questions:

- How do Observer scores compare with target scores?
- Are there domains that Observers found particularly difficult to score?
- Are there videos that Observers found particularly difficult to score?

The Observation Engine compares the scores entered by each Observer with the target scores provided for videos included in the study. In this case, Observer agreement is a measure of what percent of all scores match the target scores exactly. Percent discrepant is a measure of the percentage of all scores that disagree with the target score by two or more points.

#### READING THE FIGURES

The set of bars display the rate of agreement with the target score along with the proportion of scores that disagree with the target score. In this case, Figure 2 is calculated by domain and Figure 3 is calculated by video. **Note that only Observers who submitted scores for all videos in the scoring study are included in these figures.**

The thick blue bar in the center represents the percentage of the time Observers' scores matched the target scores. The thinner bars to either side of the blue bar represent the percentage of time Observers' scores disagreed with the target score. Yellow and light green bars are displayed when scores are discrepant, or two or more points higher (light green) or lower (yellow) than the target scores. The exact percentage of time that scores were discrepant in either direction is indicated in the column to the right of the graph. Teal and magenta bars are displayed when scores are only one point higher (teal) or lower (magenta) than the target scores.

#### INTERPRETATION

Figure 2 can be helpful when determining whether there were specific domains Observers had trouble scoring during an observation. The explanation of the domains may need to be clarified, or additional training on domains with low agreement may be needed.

If Figure 3 reveals that Observers have trouble agreeing on scores for a particular video, this may mean that Observers share a common misconception regarding a particular sort of lessons. For example, a group of Observers may tend to give inflated scores to lessons in which students work in cooperative groups while the rest of the group does not. These misconceptions can be addressed in training or with revisions to scoring guidelines.

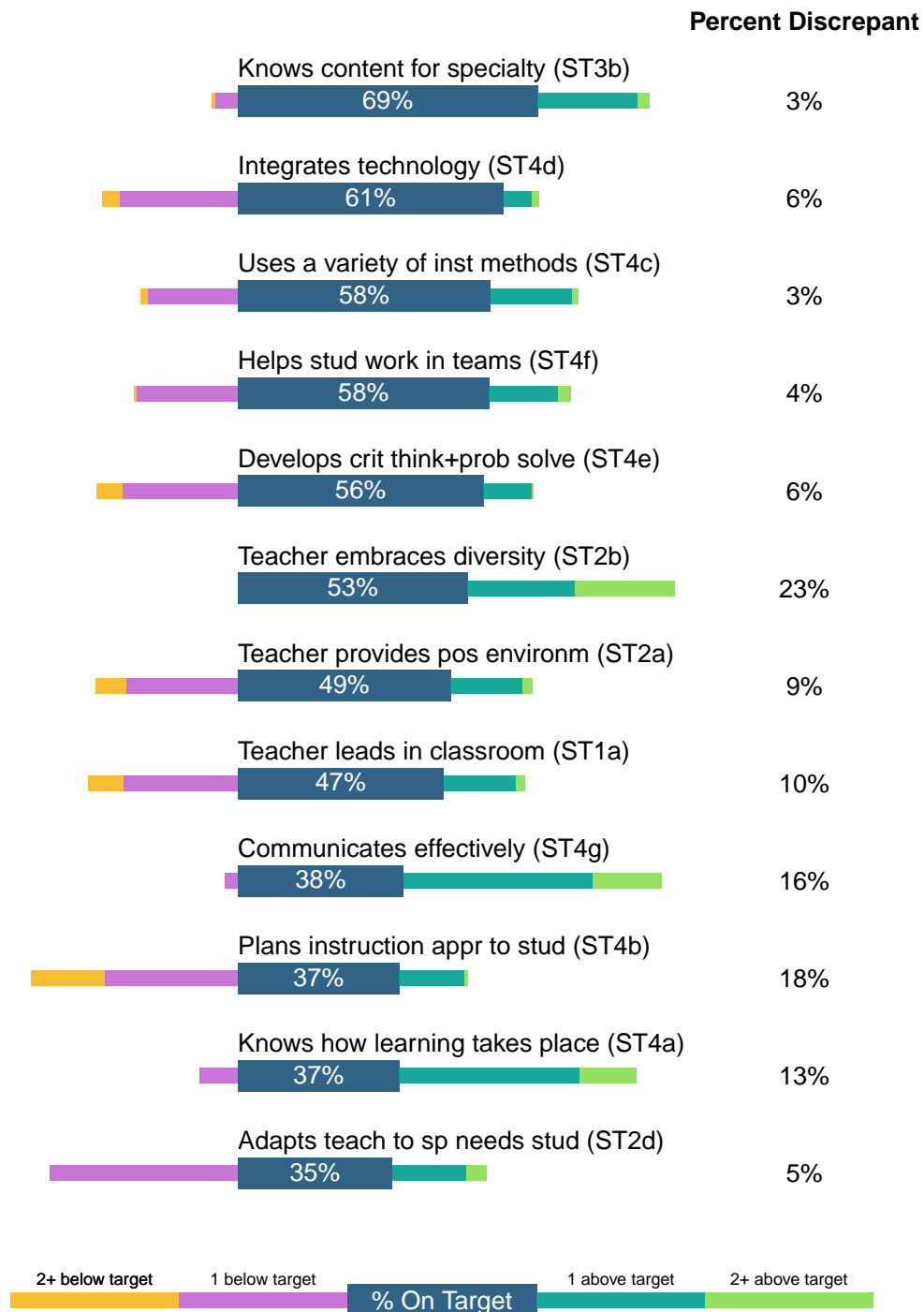


Figure 2: Observer Agreement with Target Scores by Domain



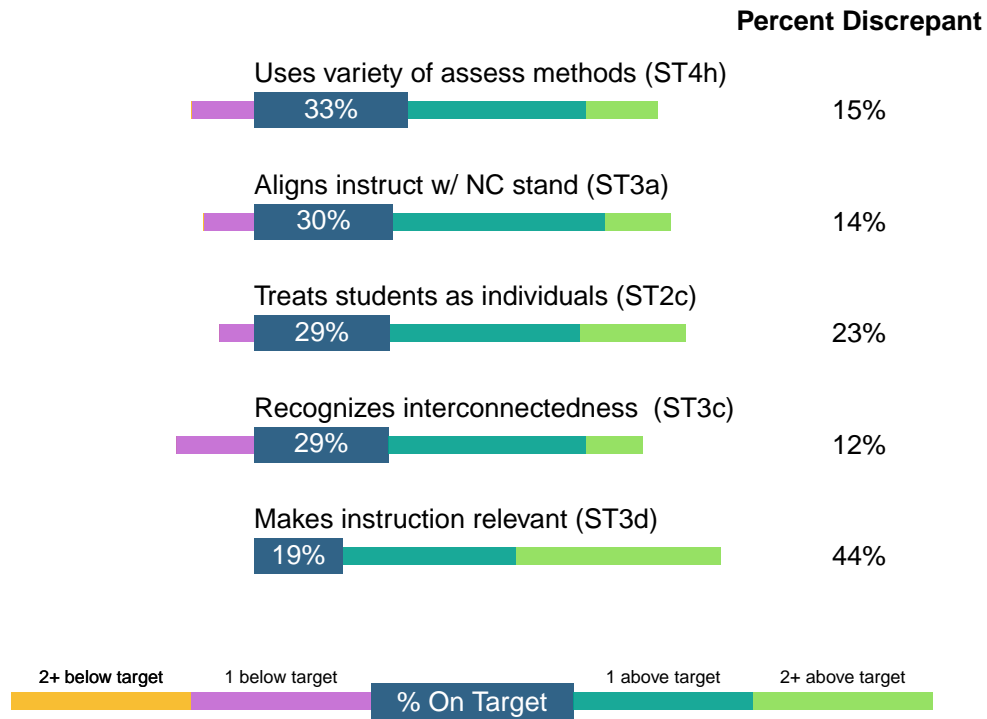


Figure 2: (Continued)

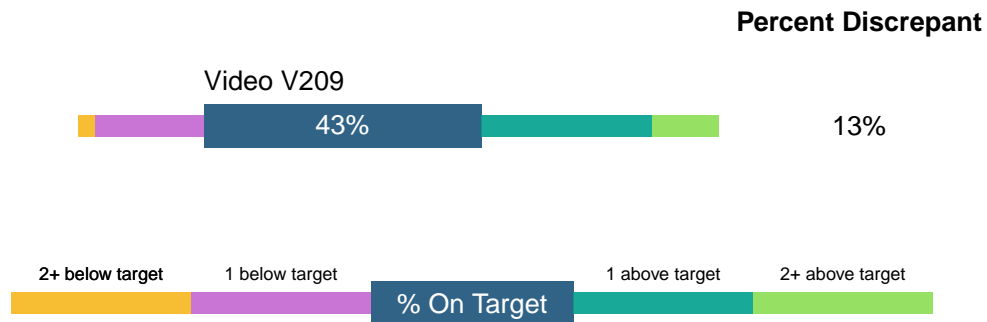


Figure 3: Observer Agreement with Target Scores by Video

## MODAL AGREEMENT

## Agreement with Modal Scores by Domain and by Video

**This section helps answer the following questions:**

- Are there domains that are particularly difficult to score?
- Are there videos that are particularly difficult to score?

### OBSERVERS' AGREEMENT WITH MODAL SCORES

In this section, the Observation Engine calculates the modal, or most common, score across all Observers for each domain and for each video. In this case, Observer agreement is a measure of what percent of all scores match the modal scores exactly. Percent discrepant is a measure of the percentage of all scores that disagree with the modal score by two or more points. To illustrate, imagine that five Observers score a video on Classroom Management. One Observer scores the video a 3, another Observer scores it a 4, and three Observers score it a 2. The modal score is 2. In this case, 60% (three out of five) of all Observers agreed with the modal score. So 60% Observer agreement would be reported for the Classroom Management domain. One out of five Observers gave a score 2 or more points away from the modal score, so a 20% rate of discrepancy (one out of five) would be reported.

### READING THE FIGURES

The set of bars display the rate of agreement with the modal score along with the proportion of scores that disagree with the modal score. In this case, Figure 4 is calculated by domain and Figure 5 is calculated by video. **Note that only Observers who submitted scores for all videos in the scoring study are included in these figures.**

The thick blue bar in the center represents the percentage of the time Observers' scores matched the modal scores. The thinner bars to either side of the blue bar represent the percentage of time Observers' scores disagreed with the modal score. Yellow and light green bars are displayed when scores are discrepant, or two or more points higher (light green) or lower (yellow) than the modal scores. The exact percentage of time that scores were discrepant in either direction is indicated in the column to the right of the graph. Teal and magenta bars are displayed when scores are only one point higher (teal) or lower (magenta) than the modal scores.

### INTERPRETATION

Figure 4 can be helpful when determining whether there were specific domains that were particularly challenging. If Observers are not certain about domain score definitions, this may result in less consistent scoring across the group. The explanation of the domains may need to be clarified, or additional training on domains with low agreement may be needed.

If Figure 5 reveals that Observers tend to disagree on scores for a particular video, this may mean that Observers are less prepared to evaluate a particular sort of lesson. For example, a subset of Observers may tend to give inflated scores to lessons in which students work in cooperative groups while the rest of the group does not. This lack of consistency can be addressed in training or with revisions to scoring guidelines.

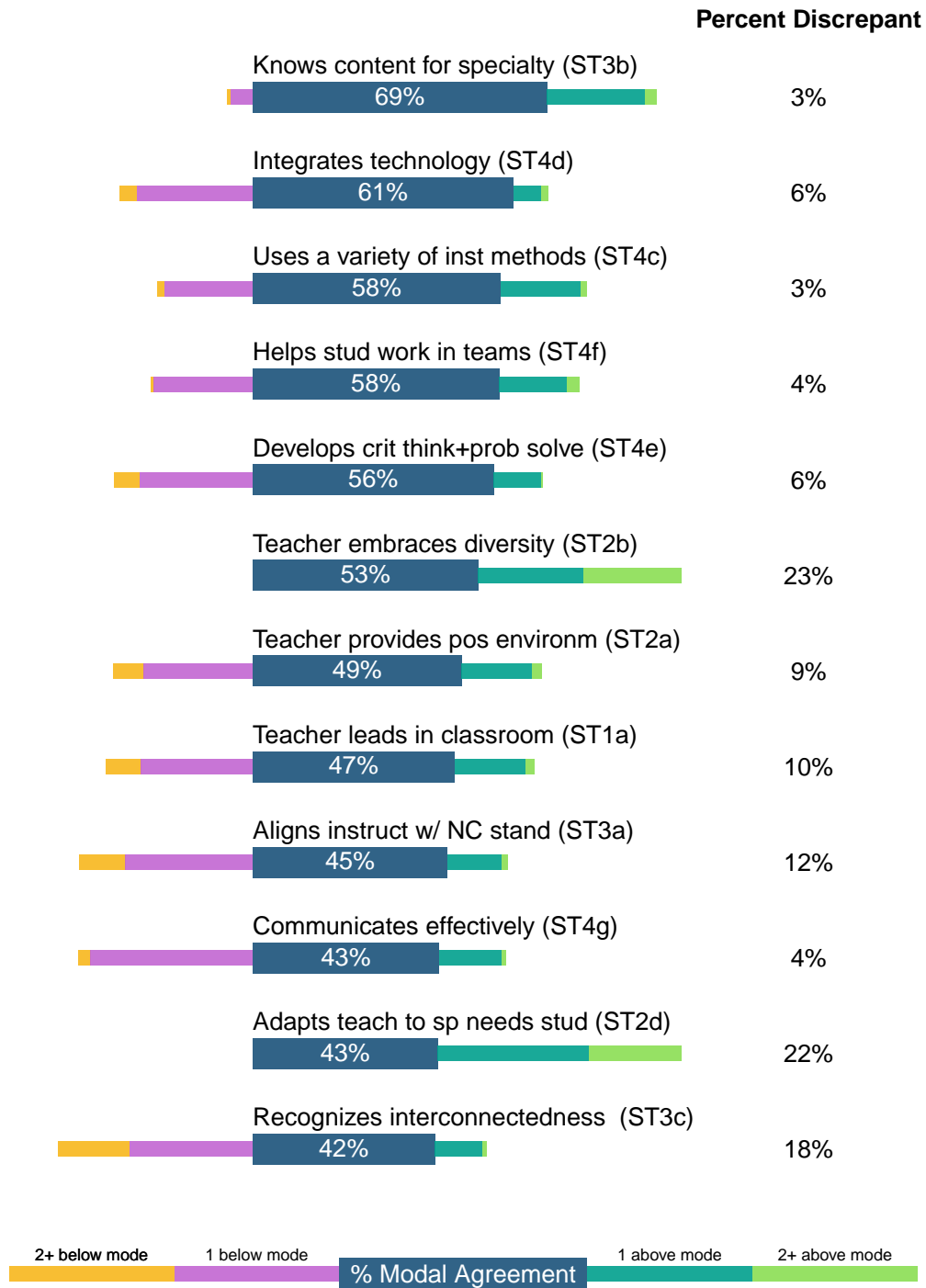


Figure 4: Observer Agreement with Modal Scores by Domain

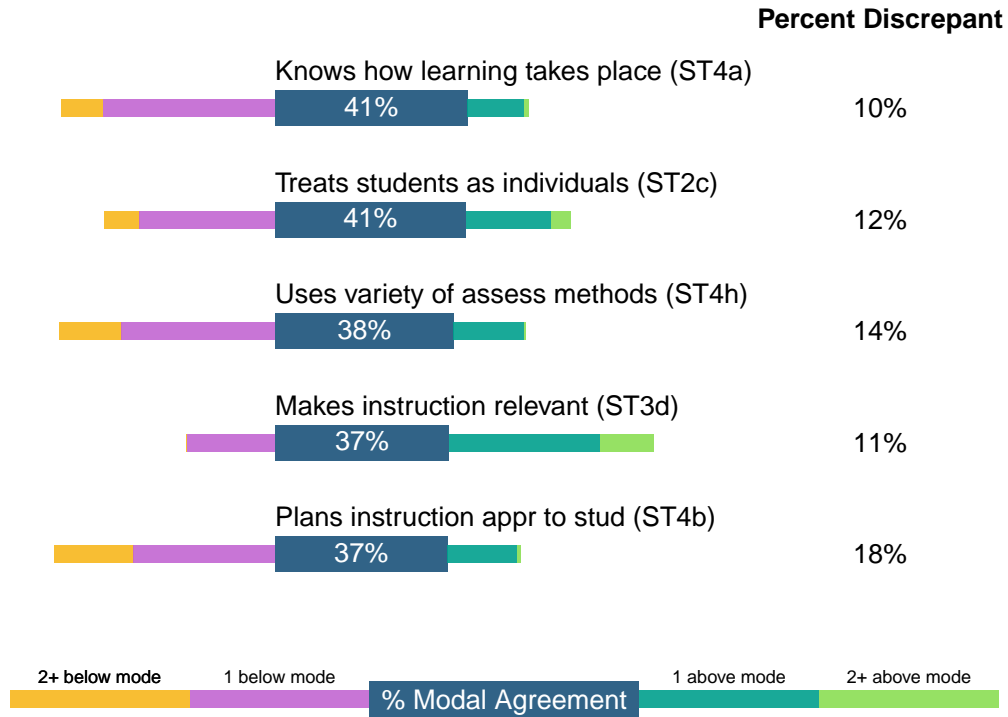


Figure 4: (Continued)

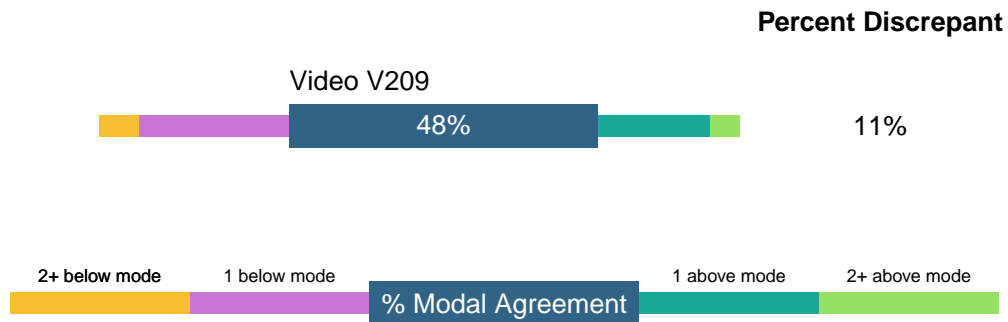


Figure 5: Observer Agreement with Modal Scores by Video