Investigating the Reliability of Classroom Observation Protocols: The Case of PLATO

M. Ken Cor
Stanford University
School of Education

April, 2011

Address correspondence to M. Ken Cor, School of Education, Stanford University, 3096 Stanford, CA 94305-3096, kcor@stanford.edu
Abstract

The reliability of the PLATO rubric is defined using Generalizability Theory and FACETS. Results from two G-Theory analyses reveal that the reliabilities of scores from individual PLATO items and the total score across all items are most strongly affected by the number of segments in which teachers’ performance is observed. For example, it is reported that a single rater should observe at least 5 segments in order to achieve a reliability coefficient greater than .80 for a score that represents the sum of scores from the separate items in the rubric on both a 12 and 7-item version of the Rubric. Reliability curves are generated to inform potential PLATO users on how measurement design characteristics affect reliability. Further, FACETS results reveal systematic effects due to occasion of measurement, segment, and rater. Scores adjusted for the specific measurement circumstances of individual observations are reported and discussed.
Investigating the Reliability of Classroom Observation Protocols: The Case of PLATO

Two sets of analysis are reported in order to establish the measurement properties of the Protocol for Language Arts Teaching Observation (PLATO) rubric. First, the expected reliability of the PLATO rubric for different measurement schemes is estimated using techniques from Generalizability Theory (See Cronbach, Gleser, Nanda, and Rajaratnam, 1972 or Brennan, 2001). Second, the measurement properties of scale scores calculated from the particular data collected in the present study are determined using FACETS (Linacre, 2010). The results from the G-Theory analysis take into account the effect of error from different raters (judges) and different observation segments on the reliability of the individual item scores as well as scores that represents the composite of the individual items that make up the PLATO rubric. The results from the FACETS analysis take into account the same sources of error as well as error from the occasion of measurement and item sampling to produce scores that are adjusted for these sources of error. While on the surface the two sets of results appear to provide two versions of the same result, they serve different purposes and are based on different assumptions described next.

The purpose of the G-theory analysis is to provide researchers interested in adopting the PLATO rubric in their research with guidelines for determining the measurement scheme required to achieve different levels of reliability. The primary finding from the G-theory analysis is a set of reliability estimates that differ depending on the measurement scheme employed. Alternatively, the FACETS analysis specifies the measurement properties of scores that have been adjusted for the specific measurement
circumstances of these particular data. These results are useful for researchers interested in using these specific data in descriptive models. The primary result from the FACETS run is a set of scaled scores and their respective standard errors.

Other than differences in the respective uses, it is also important to note how G-theory and FACETS analyses treat factors in the measurement design differently. In our G-theory analysis all tractable sources of error are considered random effects whereas in any FACETS analysis tractable sources of error are treated as fixed effects. Hence, results from the G-theory analysis become predictive of the reliability expected if for example different numbers of different raters are used to produce the data, whereas results from the FACETS run speak exclusively to the error associated with the particular raters used in the collection of these data for example.

We now present a brief description of the PLATO rubric and measurement scheme followed by a presentation of the G-theory and FACETS results respectively. The paper concludes with a brief summary of the results and a discussion of the implications of the findings.

**The PLATO Rubric and Observation Protocol**

The following is a very brief overview of the structure of the PLATO rubric. For a more in depth consideration of the theory informing its construction see Grossman et al (2010). The version of the rubric under consideration in the present study consists of 12 elements/items (see Table 1) designed to measure independent aspects of the quality of middle/high school English/Language Arts teacher instructional practices.

Table 1

*The PLATO Rubric Elements and their Descriptions*
<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intellectual Challenge</td>
<td>focuses on the nature of the task and the degree to which it represents a developmentally appropriate stretch or reach for the specific students</td>
</tr>
<tr>
<td>Text-based Instruction</td>
<td>measures the teacher’s ability to teach ELA strategies that can be used flexibly and independently</td>
</tr>
<tr>
<td>Classroom Discourse</td>
<td>measures the teacher’s ability to teach ELA strategies that can be used flexibly and independently</td>
</tr>
<tr>
<td>Guided Practice</td>
<td>measures the teacher’s ability to teach ELA strategies that can be used flexibly and independently</td>
</tr>
<tr>
<td>Purpose</td>
<td>focuses on the expressed clarity of ELA objectives, both in the short and long term</td>
</tr>
<tr>
<td>Representations of Content</td>
<td>captures the effectiveness of the teacher’s explanations and examples in addition to evidence of his or her content knowledge</td>
</tr>
<tr>
<td>Connections-Prior Knowledge</td>
<td>measures the extent to which new material is connected to students’ previous academic knowledge</td>
</tr>
<tr>
<td>Connections-Personal/Cultural Experiences</td>
<td>focuses on the extent to which new material is connected to students’ personal and cultural experiences</td>
</tr>
<tr>
<td>Models/Modeling</td>
<td>captures the availability of exemplars to guide student work</td>
</tr>
<tr>
<td>Explicit Strategy Instruction</td>
<td>measures the teacher’s ability to teach ELA strategies that can be used flexibly and independently</td>
</tr>
<tr>
<td>Accommodations for English Language Learners</td>
<td>measures the teacher’s ability to teach ELA strategies that can be used flexibly and independently</td>
</tr>
</tbody>
</table>
Classroom Environment focuses on the degree to which behavior management facilitates academic work


Next, the observation protocol used to generate the data analyzed for the present study called for up to two out of 13 raters to observe 181 teachers in up to 9 segments of instruction. For a detailed description of the characteristics of the teacher sample that was used for the analysis see Grossman, Cohen, and Brown (2011). The G-Theory results are presented next.

**Generalizability Theory Results**

Generalizability theory facilitates the investigation of how different sources of variation in the measurement design contribute to the reliability the PLATO rubric. There are two steps in a G-Theory analysis required in order to produce reliability estimates. First, estimates of the variance contributions of different sources of error in the measurement design are calculated using a G-Study. Second, the estimates from the G-study are used in Dependability Studies (D-studies) to calculate expected reliability estimates for different measurement schemes. The G-study results are presented next.

**The G-Study**

Two potential sources of error are investigated for their contribution to observed score variance over and above that which is attributable to differences in teacher ability on each of the 12 items of the rubric. First, we consider whether randomly selected raters produce similar relative rankings of the teachers for each item. Second, we assess if randomly selected segments of a teacher’s lesson produce similar relative rankings of teachers for each item.
In order to conduct the G-study, the data to be analyzed must fit a single measurement design. The measurement design describes how scores are structured in the data collection process. For example, in a study where the same two raters provide a score on a single item for all teachers, a fully crossed teacher by rater measurement design (t x r) describes all of the data. Unfortunately, the data for our study do not fit a single measurement design. Therefore, in the first step of the analysis a subset of data that fit a single measurement design is identified. Based on this process a subset of 106 out of 181 teachers observed in the first three consecutive segments by two different raters are identified for analysis within the framework of a G-Study. In order to use this data we ignore the fact that raters differ depending on the teacher and assume a teacher by rater by segment multivariate design (t' x s' x r' where the solid circles indicate the model applies to all items on the rubric). In other words we assume teachers are observed by the same two raters three separate times for each item in the rubric\(^1\). Based on this design the G-Study produces estimates of the variance contributions of the individual factors and their interactions for each item (Table 2).

Table 2

*Source Table for t' x s' x r' Design Based on Observations from 106 Teachers Over 3 Segments with 2 Raters in Each Segment*

<table>
<thead>
<tr>
<th>Source</th>
<th>intc</th>
<th>tbi</th>
<th>dis</th>
<th>gui</th>
<th>pur</th>
<th>roc</th>
<th>cpk</th>
<th>cpe</th>
<th>mo</th>
<th>esi</th>
<th>all</th>
<th>env</th>
</tr>
</thead>
</table>

\(^1\)The same data was analyzed under the assumption that different pairs of raters observe each teacher. The resulting model is a rater within teacher by segment multivariate model ((r';t') x s' where the solid circles indicate each factor is crossed with all items). The results from this design were the same as for the t' x s' x r' design suggesting our assumption has little to no effect on the results because there is little to no rater effect.
The Reliability of PLATO 8

Teacher | .17 | .45 | .20 | .17 | .08 | .08 | .15 | .22 | .13 | .22 | .13 | .19
Segment | .03 | .07 | .02 | .11 | .00 | .00 | .16 | .00 | 0   | 0   | 0   | 7
Rater   | .00 | .00 | 0   | 0   | 0   | .00 | .00 | 0   | 0   | 0   | 0   | .00
Teacher x Sgmnt | .24 | .62 | .36 | .49 | .16 | .09 | .82 | .63 | .61 | .39 | .18 | .09
Teacher x Rater | .00 | .02 | .04 | .00 | .00 | .04 | .02 | 0   | 0   | 0   | .02 | .01
Segment x Rater  | 0   | 0   | 0   | 0   | .00 | 0   | 0   | 0   | 0   | 0   | 0   | 0
Error    | .11 | .18 | .13 | .16 | .06 | .06 | .26 | .13 | .19 | .12 | .07 | .04

Note. intc = intellectual challenge, tbi = text based instruction, gui = guided instruction, pur = purpose, roc = representations of content, cpk = connections to prior knowledge, cpe = connections to personal/cultural experience, mod = models/modeling, esi = explicit strategy instruction, all = accommodations for English Language Learners, env = classroom environment.

With the exception of cpk and mod items, we see that the Teacher by Segment, the Teacher, and the Error variance contributions consistently rank first, second, and third across the 12 items based on their absolute contribution to the observed score variance.

For the cpk and mod items, the rank of the Teacher and Error variance contributions are reversed. These results suggest 1) teachers’ scores on the 12 items vary depending on the segment they are observed in, 2) there is measureable variation in teachers’ scores on the 12 items, and 3) scores from the 12 items contain undocumented error.

To make more nuanced interpretations, we consider the proportion of variance contributed to the observed score from each source across the 12 items (Table 3).

Table 3

Proportion of Total Variance for t’ x s’ x r’ Design Based on Observations from 106 Teachers Over 3 Segments with 2 Raters in Each Segment

| Item | .17 | .45 | .20 | .17 | .08 | .08 | .15 | .22 | .13 | .22 | .13 | .19
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----
| Teacher | .4  | 5   | 2   | 7   | 9   | 8   | 9   | 0   | 5   | 7   | 9   | 9
| Segment | .03 | .07 | .02 | .11 | .00 | .00 | .16 | .00 | 0   | 0   | 0   | .00
| Rater   | .00 | .00 | 0   | 0   | 0   | .00 | .00 | 0   | 0   | 0   | 0   | .00
| Teacher x Sgmnt | .24 | .62 | .36 | .49 | .16 | .09 | .82 | .63 | .61 | .39 | .18 | .09
| Teacher x Rater | .00 | .02 | .04 | .00 | .00 | .04 | .02 | 0   | 0   | 0   | .02 | .01
| Segment x Rater  | 0   | 0   | 0   | 0   | .00 | 0   | 0   | 0   | 0   | 0   | 0   | 0
| Error    | .11 | .18 | .13 | .16 | .06 | .06 | .26 | .13 | .19 | .12 | .07 | .04

2   | 9   | 3   | 9   | 6   | 2   | 3   | 3   | 6   | 2   | 0   | 6
The reliability of PLATO

Source | intc | tbi | dis | gui | pur | roc | cpk | cpe | mod | esi | all | env
--- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | ---
Teacher | .30 | .33 | .27 | .17 | .27 | .34 | .11 | .21 | .14 | .30 | .33 | .55
Segment | .06 | .05 | .02 | .11 | .01 | .00 | .11 | 0  | .00 | .00 | .01 | .9
Rater   | .00 | .00 | 0  | 0  | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .9
Teacher x | .42 | .46 | .48 | .49 | .48 | .38 | .56 | .62 | .64 | .53 | .43 | .26
Sgmnt   | 8  | 3  | 9  | 8  | 6  | 6  | 8  | 9  | 9  | 1  | 8  | 6
Teacher x Rater | .01 | .03 | .04 | .02 | .01 | .02 | .02 | .00 | .00 | .05 | .02
Segment x | .00 | 2  | 1  | 0  | .00 | 3  | 8  | 9  | 0  | 0  | 4  | 8
Rater   | 0  | 0  | 0  | 1  | 4  | 0  | 0  | 0  | 0  | 0  | 0  | 0

Based on the proportions, we see the ratio of observed score variance attributable to differences in teacher ability to total variance ranges from a low of .110 for the cpk item to a high of .556 for the env item. These results suggest that all items appear to be measuring some distinguishable aspect of teacher performance but that they do so to varying degrees. Next, the presence of variance attributable to Segment suggests that the average scores for each of the three segments (average score within each segment produced from all raters and all teachers) vary somewhat depending on the segment being observed. The gui and cpk items are most susceptible to this type of variation with proportions of total variance equal to ed for different measurement schemes. For example, it is easy to imagine how the reliability of average scores for teachers on each item of the PLATO rubric should increase as a function number of raters observing and number of segments observed. D-studies facilitate the modeling of how changes in the number of levels of the different factors affect the reliability of the rubric scores.
In our first D-study, we use a single rater by six segment design to investigate the reliability of each item in the rubric, the reliability of the scores for each dimension, and the reliability of scores produced for the entire rubric (Table 5).

Table 5

*Generalizability Coefficients Based on a 1 Rater 6 Segment D-Study*

<table>
<thead>
<tr>
<th>Item</th>
<th>G-Coeff</th>
<th>Composite G-Coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>intc</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>tbi</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>dis</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>gui</td>
<td>0.54</td>
</tr>
<tr>
<td>Factor 2</td>
<td>pur</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>roc</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>cpk</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>cpe</td>
<td>0.60</td>
</tr>
<tr>
<td>Factor 3</td>
<td>mod</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>esi</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>0.69</td>
</tr>
<tr>
<td>Factor 4</td>
<td>env</td>
<td>0.86</td>
</tr>
<tr>
<td>Entire Rubric</td>
<td></td>
<td>0.81</td>
</tr>
</tbody>
</table>

The numbers in Table 5 can be interpreted in a similar way as classical reliability coefficients such as Cronbach’s alpha. For example, the G-coefficient of .86 reported for the *env* item in column one suggests that the internal consistency of a test that consists of
The Reliability of PLATO

6 items—one for each rater/segment observation—is quite high. In other words, teachers appear to be ranked consistently across the 6 separate rater/segment observations. Based on this interpretation and the numbers in column one, we see that \( env \) chart shown in Figure 4 there is an effect due to occasion.

![Characteristics of Occasion Effects as Determined by FACETS](image)

**Figure 4.** Characteristics of Occasion Effects as Determined by FACETS

Occasion 8 appears to be the most difficult. As a result, teachers who are rated up to 8 times will end up with systematically lower scores than teachers rated only seven times. That being said, the rest of the occasions are similarly difficult and as a result would not likely produce systematically different scores for teachers as indicated by the
overlapping bubbles because occasions whose bubbles overlap are not significantly
different in their severity.

The bubble chart also reveals that occasions fit the model well as indicated by the
bubbles clustering close to the center line. In other words, occasions appear to be
functioning in a similar way. Finally, the bubble chart reveals that occasions vary in the
precision with which their overall effect is estimated. For example, occasion 8 produces
the largest standard error (largest bubble) which is a direct result of the fact that this
occasion is only ever observed on a small number of teachers. The Segment facet is
considered next.

**Segments**

Teachers are observed up to four times within any given lesson. Not all teachers
are observed four separate times and therefore teachers scores might depend on the
number of times they were observed. According to the bubble chart shown in Figure 5
there is an effect due to segment.
Figure 5. Characteristics of Segments Effects as Determined by FACETS

Segment four is the most difficult and teachers who are rated up to four times will end up with systematically lower scores than teachers rated only three times. Further, the chart shows that the second segment is the least difficult. Next, we see that the model describes the Segment facet well as indicated by the bubbles clustering close to the centerline. Finally, the fourth segment is measured with the least degree of precision.

The Rater facet is considered next.

Raters

Teachers could be rated by up to two out of 13 total raters. Therefore, not all raters provide judgments for each teacher. As a result, the score a teacher gets could
depend on which raters they are rated by. According to the bubble chart shown in Figure 7, we see that there does appear to be an effect due to rater.

Figure 7. Characteristics of Rater Effects as Determined by FACETS

For example, Rater 4 appears to be the most severe. Teachers who are only ever rated by Rater 4 will end up with systematically lower scores than teachers rated by Rater 3 for example. That being said, it does appear that most raters are similarly severe and as a result would not likely produce systematically different scores for teachers as indicated by the overlapping bubbles.

Figure 7 reveals that the model describes raters scoring behavior well as indicated by the bubbles clustering close to the center line. In other words, raters appear to be using the rubric in a similar way despite the fact that they differ in their severity. Finally,
raters vary in the precision with which their overall effect is estimated. For example, rater 12 produces the largest standard error which is a direct result of the fact that this rater rated only a small number of teachers. Prior to moving on to the Item facet one more piece of useful information from the FACETS output for the Rater facet is considered.

For the Rater facet a table of exact agreement showing the percentage of times each rater is in exact agreement with all other raters is produced (Table 7).

Table 8

*Percentage of Instances Raters are in Exact Agreement*

<table>
<thead>
<tr>
<th>Rater</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
</tr>
<tr>
<td>2</td>
<td>73</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
</tr>
<tr>
<td>6</td>
<td>84</td>
</tr>
<tr>
<td>7</td>
<td>78</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>77</td>
</tr>
<tr>
<td>10</td>
<td>85</td>
</tr>
<tr>
<td>11</td>
<td>81</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>13</td>
<td>71</td>
</tr>
</tbody>
</table>
From these results we can identify problematic raters – raters that do not agree as much as other raters. Clearly, Rater 12 had the hardest time agreeing with other raters contributing to the large standard error we see in the bubble chart shown in Figure 7. Next, while Raters 1 and 13 are the next most disagreeable, they are nowhere near as problematic as Rater 12. These results are beginning to raise flags about whether the data from Rater 12 are trustworthy. We now move on to the Item facet.

**Items.** Because all teachers are assessed on all items, there is no main effect due to item that can impact teachers overall scores. However, the FACETS output for the Item facet is useful in order to get a sense for the difficulty profile of the items. Figure 8 shows the bubble chart that results from the FACETS run.

*Figure 8. Characteristics of Item Effects as Determined by FACETS*
The bubble chart reveals that the _all_ (Accommodations for English Language Learners) is the most difficult item while the _env_ item is the easiest. However, in general there is a distribution of difficulty across the items. Next, all items are described well by the model but some items are measured with less precision. While the differences in precision are usually mostly attributed to fewer instances of measurement, for the Item facet all items are measured the same number of times. Therefore, the differences in precision across items are due to lower consistency across raters and segments for different items. For example, the _roc_ item is likely harder for raters to agree on than the other items. Teacher scale scores are presented next.

**Teacher Scale Scores.** The primary result of the facets run is the adjusted teacher scores from the rubric. The following results are estimates of teacher scores averaged across all occasions, segments, raters, and items adjusted for the particular measurement circumstance that gives rise to the observations.
Figure 9. Teacher Scale Scores

The bubble chart shows that there is a range of performance for the 181 teachers. There is also a varying degree precision for each score. Most teachers scores have similar standard errors (bubble size) however, a few are measured with less precision. Further, the model appears to describe some teachers poorly. Taken together the evidence suggests that the PLATO instrument works well at describing the performance of most if not all teachers. There are some teachers whose performance is less well measured by PLATO. These teachers should be considered further to determine if they have any distinguishing characteristics that might explain why the instrument was less precise at measuring their performance. A brief concluding discussion is presented next.

Conclusion
In the present paper we have laid out two sets of analyses that define the measurement properties of the PLATO rubric. First, our G-theory results showed that the PLATO rubric is predicted to be highly reliable under many different measurement schemes. The most important factor in determining the reliability of the rubric was revealed to be the Segment factor. Results showed that the teachers’ performance varies substantially depending on the segment being observed. As a result, it is estimated that at least 5 segments are required to achieve an overall reliability index greater than .80. This result applied to both a 12 and 7-item version of the rubric. Further, the Rater factor was shown to have little impact on the overall reliability of the rubric with as few as one rater and five segments being required to produce a reliability index greater than .80. While this result is promising from a standpoint of inter-rater reliability, it must be interpreted within the context of the training that was used to endure raters knew how to use the rubric. As a result, in order to expect the same high inter-rater reliability found in these results, the strict training requirements should be employed. Taken together, searchers interested in adopting the PLATO rubric in their studies can use the G-theory results reported here to help design a measurement scheme likely to produce highly reliable scores.

Next, our Facets results revealed Occasion, Segment, and Rater effects. Teachers who are observed in more lessons/occasions will likely have lower average scores than teachers observed fewer times because the eighth segment is the most difficult. Similarly, teachers observed in more segments will likely have lower scores because the fourth segment is the most difficult. Finally, teachers observed more frequently by Rater 4 or 8 will likely have lower scores than teachers observed by other raters. These results
suggest that scores adjusted for the specific measurement context of a given observation are more comparable and likely stronger indicators of teacher performance. Researchers interested in using these data in predictive models are advised to use the adjusted scale scores because they adjust for the measurement design features that give rise to systematically higher or lower scores.
References


