



Measuring the Power of Learning.™

**Research Report**  
ETS RR-18-11

# Agreement of Teachers on Evaluating Assessments of Learning Progressions in English Language Arts and Mathematics

---

Peter van Rijn

Edith Aurora Graf

Meirav Arieli-Attali

Yi Song

December 2018

Discover this journal online at  
**Wiley Online Library**  
wileyonlinelibrary.com

# ETS Research Report Series

---

## EIGNOR EXECUTIVE EDITOR

James Carlson  
*Principal Psychometrician*

## ASSOCIATE EDITORS

Beata Beigman Klebanov  
*Senior Research Scientist*

Heather Buzick  
*Senior Research Scientist*

Brent Bridgeman  
*Distinguished Presidential Appointee*

Keelan Evanini  
*Research Director*

Marna Golub-Smith  
*Principal Psychometrician*

Shelby Haberman  
*Distinguished Research Scientist, Edusoft*

Anastassia Loukina  
*Research Scientist*

John Mazzeo  
*Distinguished Presidential Appointee*

Donald Powers  
*Principal Research Scientist*

Gautam Puhan  
*Principal Psychometrician*

John Sabatini  
*Managing Principal Research Scientist*

Elizabeth Stone  
*Research Scientist*

Rebecca Zwick  
*Distinguished Presidential Appointee*

## PRODUCTION EDITORS

Kim Fryer  
*Manager, Editing Services*

Ayleen Gontz  
*Senior Editor*

---

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

## RESEARCH REPORT

# Agreement of Teachers on Evaluating Assessments of Learning Progressions in English Language Arts and Mathematics

Peter van Rijn,<sup>1</sup> Edith Aurora Graf,<sup>2</sup> Meirav Arieli-Attali,<sup>3</sup> & Yi Song<sup>2</sup>

1 ETS Global, Amsterdam, Netherlands

2 Educational Testing Service, Princeton, NJ

3 Fordham University, New York, NY

In this study, we explored the extent to which teachers agree on the ordering and separation of levels of two different learning progressions (LPs) in English language arts (ELA) and mathematics. In a panel meeting akin to a standard-setting procedure, we asked teachers to link the items and responses of summative educational assessments to LP levels. We evaluated four types of agreement among teachers, with test developers, and with trained raters. Although the results were quite mixed, agreement on the LP levels of mathematics items was generally better than that for ELA tasks. The agreement on the LP levels of ELA sample responses was generally better than that for mathematics. Implications of the results are discussed.

**Keywords** Learning progressions; summative assessment; mathematics; ELA; rater agreement; teachers

doi:10.1002/ets2.12199

A learning progression (LP) is a characterization of how student thinking evolves over time with respect to a particular topic or practice. It consists of ordered levels, which are qualitative descriptions of observable student behaviors as thinking develops from intuitive (and sometimes problematic) ideas to conceptions that are increasingly advanced, nuanced, and precise (see Confrey, Maloney, Nguyen, Mojica, & Myers, 2009; Corcoran, Mosher, & Rogat, 2009; Deane, Sabatini, & O'Reilly, 2012; National Research Council, 2006, 2007; Simon, 1995; Smith, Wiser, Anderson, Krajcik, & Coppola, 2007). Research on the development and validation of LPs is now widespread in the educational sciences (e.g., Alonzo & Gottwals, 2012; Black, Wilson, & Yao, 2011; Briggs & Peck, 2015; Graf & van Rijn, 2016; Steedle & Shavelson, 2009; van Rijn, Graf, & Deane, 2014; West et al., 2012; Wilmot, Schoenfeld, Wilson, Champney, & Zahner, 2011; Wilson, 2009), but there is relatively little research on how an important group of end users (teachers) interprets them for the purpose of assessing and guiding students (Furtak & Heredia, 2014). In the context of assessment, it cannot be assumed that all teachers (or all researchers) will interpret an LP in the same way, or in the way intended by the developers of the progression, and if they do not, the same student performance may be assigned to different levels of a progression by different users. This performance starts at the item response level, but misinterpretation can occur at different levels of aggregation (e.g., via some psychometric model to assign students to levels). Because an LP may be used as a guide for instruction, the assignment of different levels to the same performance implies that a student may receive different instructional activities depending on who assesses his or her standing on the progression. Such disagreement might occur because the progression itself is ambiguous, because its users have not been sufficiently trained in its intended interpretation, or because the tasks that elicit student performances are not clearly linked to the progression.

In this study, we explored the extent to which teachers agree on the ordering and separation of levels of two different LPs. Our focus is on the context of educational assessment. In this context, we can argue that a formative hypothesis generated from a summative educational assessment would be placement at a level in an LP. Our goal is to investigate how such formative hypotheses are interpreted by groups of teachers.

The capacity to generate formative hypotheses has been an explicit objective of the *CBAL*<sup>®</sup> learning and assessment tool: "Along with evaluating overall performance, the intention is that the summative assessment provides one or more formative hypotheses about student standing in an LP, which teachers should confirm or refute through follow-up classroom assessment" (Bennett, 2011, p. 14). Hence this research is related to *CBAL* projects on LPs. In this study, we convened two

*Corresponding author:* P. van Rijn, E-mail: pvanrijn@ets.org

teacher panels to evaluate the link between LPs in both English language arts (ELA) and mathematics using tasks, items, and sample responses from summative educational assessments.

As noted, an LP is a theory of how student understanding about a concept or topic develops. Our goals are to evaluate the validity of the theory and the extent to which assessment tasks provide evidence that a student is at a particular level. Deane *et al.* (2012) characterized an LP as follows:

- An LP is a description of qualitative change in a student's level of sophistication for a key concept, process, strategy, practice, or habit of mind.
- Change in student standing on such a progression may be due to a variety of factors, including maturation and instruction.
- Each progression is presumed to be modal, that is, to hold for most, but not all, students.
- It is provisional, subject to empirical verification and theoretical challenge.

When it comes to assigning LP levels, a distinction can be made between the assessment itself and the performance on this assessment. That is, an assessment (or its items) can target multiple LP levels, but the performance on the assessment is typically classified into a single LP level (with uncertainty). Furthermore, LP levels can be assigned to performance at different grain sizes of the assessment:

- *assessment*: An LP level is assigned to the complete set of items composing an assessment.
- *task*: An LP level is assigned to a subset of items, where multiple tasks are used in an assessment.
- *item*: An LP level is assigned to a single item, where a correct answer is indicative of a particular level. Note, however, that an incorrect answer can be any level below the LP level assigned to the correct answer.
- *response*: Different responses or response options can be assigned different levels of an LP (e.g., partial credit scoring can be used when there are more than two score categories and each score category is directly mapped onto levels in the LP). Also, for selected-response items, different response options can be linked to different levels in an LP but also to the same level (e.g., different misconceptions can be common in a certain lower level).

Note that assigning an LP level for a full assessment (or task) is usually done by some aggregation method or model. However, for items and responses, this is more typically done by raters. For the latter, this is the place where we want to make sure raters are in sync, and this is also our focus.

In the present study, we asked teachers to link the items of summative educational assessments to particular LPs. This was done for an LP in ELA and for one in mathematics. We used a process akin to standard-setting and alignment procedures (Cizek & Bunch, 2007). In this procedure, two panels consisting of ELA and mathematics teachers were convened, trained in the LP, and asked to associate levels in the LPs with items and tasks. In addition, we asked them to link sample responses to a subset of items to levels in the LP. Finally, the ELA teachers were asked to rate the difficulty of the texts that were used in the ELA assessments (e.g., an item can be at a low LP level but the associated text can be difficult, and vice versa). In a similar fashion, the mathematics teachers were asked to rate the computational complexity of the mathematics items (e.g., an item can be at a high LP level but the computations needed to arrive at the solution can be easy, and vice versa).

If formative hypotheses about student placement into LP levels are to be meaningful, certain psychometric criteria must be met. Since our proposed procedure is in many ways analogous to standard setting, we followed the guidelines on standard setting from the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) to determine which psychometric considerations were relevant. The standards recommended providing a description of the procedure, including training and instruction, decision process, and level of agreement among raters (see also Cizek, Bunch, & Koons, 2004). The psychometric focus in this study was therefore on agreement, and we can distinguish different types to answer our four main research questions.

The first research question concerns the evaluation of the extent to which the teachers agree among themselves on the LP levels of items. The second research question to address is the extent to which the teacher ratings of the LP levels of items agree with the LP levels that test developers<sup>1</sup> assigned to the items. This agreement is to be evaluated on an individual basis, but also the modal teacher rating will be used. It is important to evaluate both types of agreement, because high agreement among teachers will not guarantee high agreement with the test developer, and vice versa. In addition, the

higher the agreement among teachers is, the more confidence we might have that the LP level represented by the group statistic is a meaningful one, in the sense of being shared across individuals.

The third research question deals with the agreement among teachers on LP levels assigned to actual responses. Finally, our fourth research question is concerned with the agreement about LP level ratings of sample student responses of teachers to the ratings of a trained rater. Again, this agreement will be evaluated on individual and aggregate bases. These two types of agreement are important, because agreement on which LP levels the items address does not guarantee agreement on LP levels assigned to actual responses. In sum, we argue that higher agreement for all four types enhances the validity of the formative hypotheses about student placement at an LP level.

To our knowledge, this is one of the few studies in which teachers were asked to evaluate LPs (e.g., Furtak, 2012; Furtak & Heredia, 2014). Hence the purpose of this study is largely exploratory. We did expect, however, that teachers would not be familiar with the notion of an LP and how it relates to educational assessment. Consequently, our expectations about the level of agreement for each of the types described earlier were moderate. Nevertheless, we think that the results of this study provide useful information for follow-up studies regarding what can be expected in terms of agreement (and what sample sizes should be used).

The report is organized as follows. We start out with a description of the two LPs in ELA and mathematics. Then, we describe the teacher panel study for ELA, followed by the teacher panel study for mathematics. In the ELA study, we describe the statistical methodology in considerable detail, but this description is not repeated for the mathematics study because it is largely the same. After the results for each study, we provide a brief summary of the overall results. The report ends with a general discussion.

## Learning Progressions

The ELA LP under scrutiny consists of a set of related argumentation LPs. These are described in more detail by Song, Deane, Graf, and van Rijn (2013) and Deane and Song (2014). The development of argumentation in ELA is specified with respect to the following four LPs: (a) *appeal building* (understanding the context and audience), (b) *taking a position* (considering different perspectives), (c) *reasons and evidence* (building logical arguments), and (d) *framing a case* (structuring and presenting arguments). The levels in the particular strand “taking a position” are shown in Table 1 (Deane & Song, 2014). Although argumentation spans these four LPs, it is assumed that the levels are aligned across the LPs. That is, students would be expected to perform at a similar level across the four strands. A selection of tasks associated with two scenarios is used in this study, where the scenario provides a context and motivation for the tasks. A feature of these scenarios is that all tasks are exactly parallel in terms of the number of items and the item types, but the scenario topics are different. In addition, even though the full set of LP levels has a wide grade span, the selection of tasks is mostly suited for middle school students. Hence we use as reference a number of results on empirical recovery of this LP (e.g., task difficulty) obtained with a sample of seventh-, eighth-, and ninth-grade students (van Rijn et al., 2014; van Rijn & Yan-Koo, 2016).

For mathematics, we used a linear function LP for middle school students, which is described in further detail by Arieli-Attali, Wylie, and Bauer (2012) and Graf and Arieli-Attali (2015). The levels in this LP are shown in Table 2. The focus in this study is on the first four levels, because the fifth level is not addressed in the assessments that are used. We

**Table 1** English Language Arts: Levels of Argumentation Learning Progression “Taking a Position”

Level	Description
5	Can use others’ arguments to develop one’s own understanding and then frame one’s own position in terms that exploit the current “state of discussion.”
4	Successfully analyzes unstated assumptions, biases, and other subjective elements in a text and can use that to develop one’s own position more clearly.
3	Understands and expresses positions clearly, capturing their relationships both to similar and contrasting points of view.
2	Understands and expresses positions in writing with reasonable attention to what one knows and some ability to focus on what is important in the domain.
1	Understands the idea of taking a side in an argument and accepting or rejecting another person’s statements as true or false based on how well one thinks it fits the facts.

**Table 2** Mathematics: Levels of Learning Progression “Linear Functions”

Level	Short description	Name
5	Nonlinear functions	Changing change
4	More than one linear function	Comparing change
3	Linear function ( $y = ax + b$ )	Constant change
2	Coordinate plane ( $x,y$ )	Mutual change
1	Separate representations (numeric, spatial, symbolic)	One-dimensional change

make use of two similar scenario-based tasks, where one task focuses on the first three levels and the other on Levels 2, 3, and 4. A study on empirical recovery of this LP is described by Graf and van Rijn (2016), and we will use some of their results for reference (e.g., item score frequencies).

## Study 1: English Language Arts

### Methods

#### Panelists

Ten middle and high school ELA teachers were recruited from districts in the states of New Jersey and Pennsylvania. The panelists had between 2 and 25 years of teaching experience ( $M = 13.3$ ,  $SD = 7.8$ ) and were contacted via e-mail announcements and fliers. Two of the panelists had prior experience with CBAL ELA.

#### Materials

The following materials were prepared and distributed in advance: (a) an excerpt from Song *et al.* (2013) that summarized the argumentation LP as well as background research; (b) a detailed table including the complete descriptions of each level of the argumentation LP; and (c) an abbreviated table summarizing the contents of the detailed table. During the meeting, panelists provided multiple ratings on two computer-based ELA task sets, as described later. A Web-based review page with links to the tasks was made accessible. We only used a selection of tasks from the two CBAL ELA assessments, because the complete task sets address both summarization and argumentation LPs. The scenarios and tasks are described as follows.

1. *Ban Ads*. In this scenario, students are instructed to evaluate and analyze arguments and present their views in an essay. The scenario considers three articles about whether the United States should ban advertising for children. In the first selected task, students are asked to identify and explain problems in the reasoning and use of evidence in a given letter that supports such advertising. In the second task, students are asked to decide whether each of 10 statements gives a reason to ban or allow ads for children. The third task considers six claims followed by evidence, and students are asked to decide whether the evidence supports or weakens the claim (or does neither). The fourth task is the culminating essay, in which students are to present their views on the matter by stating their positions and to support their views with reasons and evidence.
2. *Junk Food*. As noted, the tasks in this scenario are highly similar to those in Ban Ads. The only difference is that the issue concerns selling junk food in school cafeterias.

All ratings were made on paper. Two rating forms were created: one for each of Ban Ads and Junk Food. Each form consisted of screen prints of the texts, tasks, and five sample student responses to the culminating essay task. The screen print for each text was immediately followed by the first type of prompt—“What is the difficulty level of the [...] text for 8th-grade students?”—and a set of five boxes, labeled with “Easy,” “Easy/Medium,” “Medium,” “Medium/Hard,” and “Hard.” The screen print of each task was followed by the second type of prompt—for example, “Consider the Reasons and Evidence LP. At which level in this progression would you place this item?”—and a set of five boxes labeled with Levels 1 through 5. The third type of rating considered sample student responses. The response was followed by this question: “This item assesses all the four argumentation skills: Appeal Building, Taking a Position, Reasons and Evidence, and Framing a Case. Please consider all these aspects, and give a holistic evaluation of the level of this student’s essay. At which level

in the argumentation LPs would you place this student’s essay?” The question was followed by a set of five boxes labeled with Levels 1 through 5. An example of the rating form can be found in Appendix A.

### Procedure

The advance reading materials were e-mailed prior to the meeting; panelists were asked to review the materials, contact us with any questions (none were received), and bring the materials to the meeting.

The meeting took place on Saturday, November 2, 2013, from 9:30 a.m. to 12:00 p.m. We began with a presentation in which we provided a basic definition of an LP. We discussed the three different types of ratings panelists would be providing with respect to the texts, tasks, and student responses to the tasks. The presentation included a review of the argumentation LPs (appeal building, taking a position, reasons and evidence, and framing a case). The presentation concluded with several sample items and two sample student essays (none of which were from Ban Ads or Junk Food). When assigning ratings to essays, panelists were asked to provide a holistic evaluation and to ignore surface-level errors, including grammar, spelling, and punctuation.

Panelists brought their own laptops to the session; wireless access was provided so that they could connect to the CBAL review page and view the tasks in their browsers. Following the presentation, panelists reviewed the Ban Ads task on their computers; in addition to reviewing the items, they were encouraged to interact with the task as a student might, to experiment with entering responses, and so on. Once they had finished reviewing, they provided ratings on the texts, tasks, and sample student responses on the corresponding paper forms by checking the appropriate boxes (see “Materials”). The screen prints on the forms served as reminders of the items they had reviewed on the computer. The panelists then reviewed the Junk Food task on their computers and provided ratings on the paper form.

### Statistical Analysis

As noted, the focus in our analyses is on agreement (Gwet, 2014). Because it is assumed that the LP levels are ordered, we make use of agreement coefficients that take the ordinal nature of the ratings into account. However, because the levels are qualitatively different, we do not assume an interval scale (i.e., the difference between Levels 1 and 2 is not necessarily the same as the difference between Levels 3 and 4). Note that Cohen’s kappa would be a natural choice to assess agreement, but it assumes ratings are nominal. This is dealt with by using appropriate weights (i.e., ordinal; see later). Furthermore, because a number of agreement coefficients are known to show paradoxical outcomes in certain situations (e.g., Li, 2016), we compute multiple coefficients to prevent making inferences on such an outcome. Most of the coefficients that we consider have the following structure:

$$\frac{p_a - p_e}{1 - p_e}, \quad (1)$$

where  $p_a$  is the proportion observed agreement among raters and  $p_e$  is the proportion chance agreement. Most differences between coefficients arise from a different definition of  $p_e$ .

In case of two raters, Gwet (2014) recommended the following agreement statistics: (a) Cohen’s kappa (Cohen, 1960), (b) Scott’s pi (Scott, 1955), (c) Brennan–Prediger coefficient (Brennan & Prediger, 1981), (d) Krippendorff’s alpha (Krippendorff, 2004), and (e) Gwet’s AC1 (Gwet, 2014). Gwet (2014) noted that the first three agreement statistics can show paradoxical outcomes (i.e., low values) in certain situations. This happens mostly when the marginal distributions are different (e.g., Karelitz & Budescu, 2013; Zwirk, 1988). Building on work by Janson and Olsson (2004) and Berry and Mielke (1988), Gwet (2014) recommended the following agreement statistics if there are three or more raters: (a) Conger’s generalized kappa (Conger, 1980), (b) Brennan–Prediger coefficient (Brennan & Prediger, 1981), (c) Fleiss’s generalized kappa (Fleiss, 1971), (d) Krippendorff’s alpha (Krippendorff, 2004), and (e) Gwet’s AC2 (Gwet, 2014). According to Gwet (2014), the first, third, and fourth of the preceding agreement statistics can show paradoxical outcomes. Each agreement statistic can be computed using a weight matrix, and several types of weights can be used (e.g., linear, quadratic). We make use of the weights for ordinal ratings, which are given by (Gwet, 2014, Equation 3.5.1)

$$w_{jk} = \begin{cases} 1 - m_{jk}/m_{1q}, & \text{if } j \neq k, \\ 1, & \text{if } j = k, \end{cases} \quad (2)$$

**Table 3** Interpretation of Kappa Values According to Landis and Koch (1977)

Kappa	Agreement
<.00	Poor
.00–0.20	Slight
.21–.40	Fair
.41–.60	Moderate
.61–.80	Substantial
.81–1.00	Almost perfect

where  $m_{jk} = \binom{r}{2}$  with  $r = \max(j, k) - \min(j, k) + 1$ , and  $q$  is the highest category. We made use of the R code provided by Gwet (2014) to estimate all agreement statistics, their standard errors, and the 95% confidence intervals.<sup>2</sup>

To interpret the values of the estimated agreement statistics, several benchmarks can be used (see Gwet, 2014, Chapter 6). For example, according to Landis and Koch (1977), the values of (generalized) kappa can be interpreted as shown in Table 3. Although this interpretation is concerned with kappa only and is subjective (other interpretations are available, e.g., Altman, 1990; Fleiss, Levin, & Paik, 2003), we adhere to this table to make qualitative judgments about agreement.<sup>3</sup>

One of the situations in which kappas can show low values is when the marginal distributions are not equal. This also holds for Scott's pi. Zwick (1988) therefore suggested first inspecting marginal homogeneity and then computing rater agreement statistics. However, this is quite cumbersome if there are many raters and thus many marginals (e.g., with 10 raters, the number of rater pairs is 45). The number of categories and sample size can also have a strong influence on the outcomes (see, e.g., Sim & Wright, 2005). For these reasons, we not only compute several different estimates but also determine the standard errors and confidence intervals and plot distributions of agreement statistics for teacher pairs.

Because some variables we would like to compare have different numbers of categories (so that agreement statistics cannot be computed), we also create contingency tables for several combinations of variables to inspect dependencies. In addition, we compute simple rank order correlations (Spearman's  $\rho$ ). Also, we compute Goodman and Kruskal's gamma (Goodman & Kruskal, 1954) but note that it performs less well in small samples than Spearman's rank order correlation (Gans & Robertson, 1981). Finally, we make use of the concept of mutual information (MI; Brillinger, 2004). In addition to the correlations, MI is useful because it is sensitive to both linear and nonlinear statistical dependence. For discrete variables, we can compute MI by letting  $P(X_1 = j, X_2 = k) = p_{jk}$ , so that we can write

$$MI(X_1, X_2) = \sum_{j,k} p_{jk} \log \frac{p_{jk}}{p_{j+}p_{+k}}, \quad (3)$$

where the sum is over all  $j, k$  for which  $p_{jk} \neq 0$ . Note that, in principle,  $1 - \exp(-2MI(X_1, X_2))$  can be used as an  $R^2$  measure if  $P(X_1, X_2)$  is bivariate normal, because in that case, the mutual information is equal to  $-0.5 \log(1 - \rho_{X_1, X_2}^2)$  (where  $\rho$  indicates the correlation). However, the normality assumption likely does not hold in our case. If we multiply the MI in Equation 3 by 2 times the sample size, we find the familiar likelihood ratio statistic  $G^2$  for independence. However, we do not use this statistic because of our limited sample size.

## Results

One recruited teacher did not attend the meeting, so the final sample in this study consists of nine teachers.

### Text Difficulty Ratings

Table 4 shows the frequencies of the text difficulty ratings for each of the six texts in the two assessments. The mean difficulty rating of Text 1 in Ban Ads ( $M = 2.9$ ) is significantly higher than the mean of the first text of Junk Food ( $M = 1.8$ ),  $t(16) = 2.67, p = .02$ .

For the difficulty ratings of the six texts, Table 5 shows for each agreement statistic the percentages observed and chance agreement, the estimate, the standard error, and the 95% confidence interval.<sup>4</sup> It can be seen that the estimates of Conger's



**Table 4** Frequencies of Difficulty Ratings for Texts in English Language Arts Assessments

Assessment	Text	Easy	Easy/med.	Medium	Med./hard	Hard
Ban Ads	1	0	4	2	3	0
	2	2	5	2	0	0
	3	1	3	4	1	0
Junk Food	1	4	3	2	0	0
	2	3	3	3	0	0
	3	2	2	5	0	0

**Table 5** Agreement Among English Language Arts Teachers on Difficulty of Texts in English Language Arts Assessments

Statistic	% Agreement		Estimate (SE)	95% CI		Agreement
	Observed	Chance				
Conger's generalized kappa	.80	.78	.10 (.05)	-.04	.24	Slight
Brennan – Prediger coefficient	.80	.69	.36 (0.04)	.27	.46	Fair
Fleiss's generalized kappa	.80	.79	.05 (.05)	-.09	.19	Slight
Krippendorff's alpha	.80	.79	.07 (.05)	-.07	.21	Slight
Gwet's AC2	.80	.64	.45 (.05)	.31	.59	Moderate

**Table 6** Frequencies of Learning Progression Level Ratings for English Language Arts Tasks

Assessment	Task	Developer LP level	LP level rating					$P^+$ (n)
			1	2	3	4	5	
Ban Ads	1	4	0	3	5	1	0	.42 (859)
	2	1	1	0	7	1	0	.71 (816)
	3	2	0	0	8	1	0	.68 (804)
Junk Food	1	4	0	2	5	2	0	
	2	1	0	2	5	2	0	
	3	2	0	1	2	6	0	

Note. LP = learning progression.

generalized kappa, Fleiss's generalized kappa, and Krippendorff's alpha have similar (low) values. In addition, the estimate of the Brennan – Prediger coefficient is not that far from Gwet's AC2. Using the interpretations of Table 3, the agreement among the ELA teachers on text difficulty can be judged as slight to moderate.

### Learning Progression Level Ratings for English Language Arts Tasks

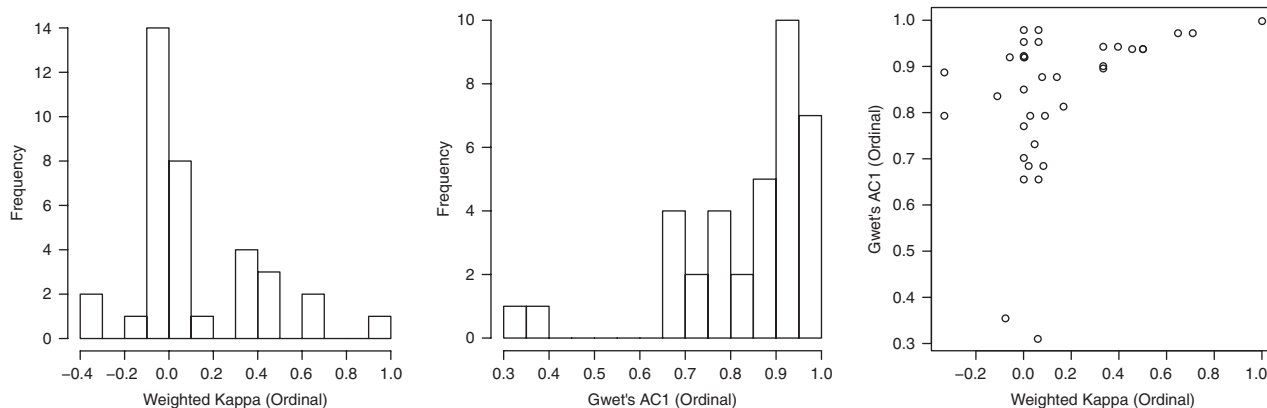
Table 6 shows the distribution of the LP level ratings for the selected tasks from the two ELA assessments. The mean level ratings for the three tasks in the two alternate forms are not significantly different. The ratings do not seem to align well with the levels used by the test developer and the difficulty of the Ban Ads tasks as indicated by the  $P^{+5}$  values from a larger study with middle school students (van Rijn & Yan-Koo, 2016).  $P^+$  values are not available for the Junk Food tasks. Finally, note that the tasks are not uniquely linked to the texts so that a comparison between text difficulty and LP level cannot easily be made.

In Table 7, the results for the agreement among ELA teachers on the LP levels of the ELA tasks are shown. The agreement among the teachers can be judged as slight to moderate (i.e., the lower bound of the confidence interval falls in the interval for slight agreement and the upper bound falls in the interval for moderate agreement). The same pattern across the different statistics is seen as for the text difficulty ratings, although the values of the Brennan – Prediger coefficient and Gwet's AC2 are somewhat higher.

Because the results are rather mixed, we also show the distributions of two agreement statistics for all 36 teacher pairs to get a sense of the variation. We used weighted kappa and Gwet's AC1, both with ordinal weights, because the results for the associated statistics with all raters are very different (e.g., Conger's generalized kappa = .08 and Gwet's AC2 = .73).

**Table 7** Agreement Among English Language Arts Teachers on Learning Progression Levels of English Language Arts Tasks

Statistic	% Agreement		Estimate ( <i>SE</i> )	95% CI		Agreement
	Observed	Chance				
Conger’s generalized kappa	.87	.86	.08 (.07)	−.10	.26	Slight
Brennan–Prediger coefficient	.87	.69	.59 (.06)	.43	.74	Moderate
Fleiss’s generalized kappa	.87	.87	.03 (.08)	−.17	.24	Slight
Krippendorff’s alpha	.87	.87	.05 (.08)	−.15	.25	Slight
Gwet’s AC2	.87	.52	.73 (.07)	.56	.90	Substantial



**Figure 1** Distributions of agreement statistics for 36 teacher pairs for learning progression level ratings of English language arts tasks: (left) weighted kappa, (middle) Gwet’s AC1, and (right) scatterplot.

Both the marginal and joint distributions are shown in Figure 1. It can be seen that weighted kappa is low if Gwet’s AC1 is low, but not vice versa. Conversely, if weighted kappa is high, Gwet’s AC1 is high, but, again, not vice versa. Furthermore, for only 2 out of 36 teacher pairs are both agreement statistics low.

Given these mixed results, we show the cross-tables of three typical pairs of ratings to study the pattern of the joint and marginal distributions. Table 8 shows the distributions of the LP level ratings of three different combinations of results: low weighted kappa – low Gwet’s AC1, low weighted kappa – high Gwet’s AC1, and high weighted kappa – high Gwet’s AC1. Clearly the LP level ratings are quite off for the first pair. However, for the second pair, the ratings are not that far off, even though weighted kappa for this pair of ratings is only .06.

Table 9 shows the agreement on the LP level of the ELA tasks of the individual teachers with the LP level assigned by the test developer. The agreement statistics are highest for Teacher 4, ranging from slight to substantial, and lowest for Teacher 2, ranging from poor to fair.

Table 10 shows the agreement statistics of the modal LP level rating of the teachers with the LP level of the test developer for the ELA tasks. The agreement ranges from poor (three out of five) to moderate.

**Learning Progression Level Ratings for English Language Arts Sample Responses**

Table 11 shows the frequencies of the LP level ratings for six sample responses to Ban Ads Task 4 and Junk Food Task 4 given by the nine teachers.

Table 12 displays the statistics for the agreement among ELA teachers on the LP levels of the responses to the two ELA tasks. In contrast to the earlier results, the statistics are now in unison: The agreement among ELA teachers can be considered substantial.

In Table 13, the statistics are shown for the agreement of the modal LP level rating of teachers with the LP level of the trained rater for the sample ELA responses. The agreement is substantial for four out of five statistics and almost perfect for Gwet’s AC2.

**Table 8** Distributions of Three Different Pairs of Ratings of English Language Arts Tasks That Lead to Different Agreement Statistics

Pair	LP level	LP level				Sum
		1	2	3	4	
Low weighted kappa – low Gwet's AC1	1	0	0	1	0	1
	2	0	0	0	4	4
	3	0	0	0	1	1
	4	0	0	0	0	0
	Sum	0	0	1	5	6
Low weighted kappa – high Gwet's AC1	1	0	0	0	0	0
	2	0	0	0	0	0
	3	0	0	0	3	3
	4	0	0	1	2	3
	Sum	0	0	1	5	6
High weighted kappa – high Gwet's AC1	1	0	0	0	0	0
	2	0	0	1	0	1
	3	0	0	4	0	4
	4	0	0	0	1	1
	Sum	0	0	5	1	6

**Table 9** Agreement Statistics for Individual English Language Arts Teachers With Learning Progression Level of Test Developer

Teacher	Cohen's kappa (SE)	Scott's pi (SE)	Brennan – Prediger coefficient (SE)	Krippendorff's alpha (SE)	Gwet's AC1 (SE)
1	-.06 (.08)	-.27 (.13)	.29 (.15)	-.17 (.13)	.45 (.15)
2	-.36 (.22)	-.54 (.17)	.11 (.26)	-.41 (.17)	.26 (.27)
3	.10 (.09)	-.24 (.26)	.11 (.30)	-.14 (.26)	.46 (.28)
4	.14 (.20)	.10 (.24)	.52 (.18)	.18 (.24)	.68 (.14)
5	-.06 (.08)	-.27 (.13)	.29 (.15)	-.17 (.13)	.45 (.15)
6	-.27 (.22)	-.44 (.14)	.17 (.14)	-.32 (.14)	.31 (.14)
7	-.12 (.24)	-.21 (.23)	.29 (.15)	-.11 (.23)	.41 (.13)
8	.00 (.00)	-.20 (.15)	.40 (.14)	-.10 (.15)	.58 (.10)
9	.00 (.15)	-.26 (.27)	.17 (.29)	-.16 (.27)	.36 (.27)

**Table 10** Agreement Statistics of Modal Learning Progression Level Rating of English Language Arts Teachers With Learning Progression Level of Test Developer for English Language Arts Tasks

Statistic	% Agreement		Estimate (SE)	95% CI		Agreement
	Observed	Chance				
Cohen's kappa	.80	.81	-.06 (.08)	-.27	.16	Poor
Scott's pi	.80	.84	-.27 (.13)	-.61	.06	Poor
Brennan – Prediger coefficient	.80	.72	.29 (.15)	-.09	.66	Fair
Krippendorff's alpha	.82	.84	-.17 (.13)	-.51	.17	Poor
Gwet's AC1	.80	.64	.45 (.15)	.07	.83	Moderate

## Summary

In sum, the agreement among ELA teachers on the difficulty level of the texts used in the ELA assessments ranged from slight to moderate. For one of the three texts, the mean rating was significantly lower in Junk Food than in Ban Ads.

With respect to our first research question, the agreement among ELA teachers on the LP level ratings of the tasks ranged from slight to substantial. For our second research question, we found that the agreement of the average LP level rating of the teachers with the LP level assigned by the test developer was qualified between poor and moderate. Although the results were quite mixed, an analysis of all 36 teacher pairs showed that only two pairs of teachers showed poor agreement for multiple statistics.

**Table 11** Frequencies of Learning Progression Level Ratings for Sample Responses to Two English Language Arts Tasks

Task	Response	Trained rater LP level	LP level rating				
			1	2	3	4	5
Ban Ads, Task 4	1	2	0	1	6	2	0
	2	1	8	1	0	0	0
	3	3	0	1	4	2	2
	4	4	0	0	6	3	0
	5	2	0	2	3	4	0
	6	3	0	1	3	5	0
Junk Food, Task 4	1	1	8	1	0	0	0
	2	3	0	1	6	2	0
	3	2	2	5	2	0	0
	4	4	0	0	3	4	2
	5	2	0	8	1	0	0
	6	2	1	3	5	0	0

Note. LP = learning progression.

**Table 12** Agreement Among English Language Arts Teachers on Learning Progression Levels of Sample Responses to English Language Arts Tasks

Statistic	% Agreement		Estimate (SE)	95% CI	Agreement
	Observed	Chance			
Conger's generalized kappa	.92	.79	.62 (.08)	.43 .80	Substantial
Brennan–Prediger coefficient	.92	.72	.71 (.05)	.61 .82	Substantial
Fleiss's generalized kappa	.92	.79	.61 (.09)	.42 .80	Substantial
Krippendorff's alpha	.92	.79	.62 (.09)	.43 .81	Substantial
Gwet's AC2	.92	.69	.74 (.05)	.62 .86	Substantial

**Table 13** Agreement of Modal Learning Progression Level Rating of English Language Arts Teachers With Learning Progression Level of Trained Rater for Sample Responses

Statistic	% Agreement		Estimate (SE)	95% CI	Agreement
	Observed	Chance			
Cohen's kappa	.94	.84	.63 (.17)	.26 1.00	Substantial
Scott's pi	.94	.85	.62 (.19)	.21 1.00	Substantial
Brennan–Prediger coefficient	.94	.72	.79 (.09)	.60 .99	Substantial
Krippendorff's alpha	.94	.85	.64 (.19)	.23 1.00	Substantial
Gwet's AC1	.94	.66	.83 (.07)	.67 .99	Almost perfect

With respect to our third research question, the agreement results for the sample responses were better and more consistent in that the agreement among teachers was uniformly substantial across the different agreement statistics. For the fourth research question, the results for the agreement of the average LP level rating with the LP level assigned by a trained rater were similar (for one statistic, the agreement was almost perfect).

## Study 2: Mathematics

### Methods

#### Panelists

The panel consisted of 10 middle and high school mathematics teachers recruited from districts in the state of New Jersey. The panelists had between 2 and 25 years of teaching experience ( $M = 10.8$ ,  $SD = 8.1$ ) and were contacted via e-mail announcements and fliers. None of the panelists had prior experience with CBAL mathematics tasks or LPs.

## Materials

The following materials were prepared and distributed to panelists in advance of the meeting: (a) an excerpt from Arieli-Attali et al. (2012) that summarized the linear functions LP as well as background research, (b) a detailed table including the complete descriptions of each level of the linear functions LP, and (c) an abbreviated table summarizing the contents of the detailed table. During the meeting, panelists provided ratings for individual items within two computer-based mathematics tasks, as described subsequently. A Web-based review page with links to the tasks had previously been established as part of a Common Core Standards alignment study (see, e.g., Tannenbaum, Baron, & Kannan, 2015). Panelists were given access to this page. Descriptions of the tasks follow:

1. *Moving Sidewalks—One Rider (MS1R)*. This task is based on a real-world scenario and assesses students' abilities to work with the slope of a line in the context of the task, understand that slope represents a rate, and translate among various representations (tables, graphs, and algebraic equations) that use slope. The MS1R task consists of 10 multipart questions, where a question includes all prompts that appear together on a single screen. The parts were aggregated into meaningful units ("items") for rating purposes, resulting in 11 items. In addition, we selected six sample student responses from each of item MS1R\_2a (Item 2, part a) and item MS1R\_06 (Item 6, parts a and b combined). At least one part from each of these two items required a constructed text response; the sample responses had been previously scored using the levels of the LP by a rater highly familiar with the Moving Sidewalks tasks. The samples from MS1R\_02a included one with a score of 1, two with a score of 2, two with a score of 3, and one with a score of 0 (indicating that the response provided insufficient evidence to be scored using any of the levels of the LP). The samples from MS1R\_06 included three with a score of 2, two with a score of 3, and one with a score of 0.
2. *Moving Sidewalks—Two Riders (MS2R)*. Although similar to MS1R, this task is more advanced and requires students to work with a linear function that has a negative slope and nonzero intercept. The MS2R task consists of 12 multipart questions, where a question includes all prompts that appear together on a single screen. The items were defined somewhat differently than for MS1R; for the sake of simplicity, each MS2R item corresponds exactly to an MS2R question. We selected six sample student responses from each of MS2R\_01ab (Item 1, parts a and b combined) and MS2R\_08ab (Item 8, parts a and b combined). As with MS1R, at least one part from each of these two items required a constructed text response; the sample responses had been previously scored using the levels of the LP by a rater highly familiar with the Moving Sidewalks tasks. The samples from MS2R\_01ab included one with a score of 1, two with a score of 2, two with a score of 3, and one with a score of 0. The samples from MS2R\_08ab included one with a score of 3 and five with a score of 0.

Although panelists first reviewed the tasks on computer, all ratings were made on paper. Two rating forms were created, one for each of MS1R and MS2R. Each form consisted of screen prints of the items, sample student responses (as described earlier), and three types of rating prompts. The screen print for each item was immediately followed by the first type of prompt—"At which level in the *Linear Functions LP* would you place this item?"—and a set of four boxes labeled with Levels 1–4. Space to provide a rationale for the selection appeared immediately below. This was followed by the second type of prompt—"For an eighth grade student, what is the expected degree of *computational complexity* involved in responding to this item?"—and a set of five boxes with the following labels: "Not complex," "Slightly complex," "Somewhat complex," "Moderately complex," and "Highly complex."

For items MS1R\_2a, MS1R\_06, MS2R\_01ab, and MS2R\_08ab, the second type of rating prompt was followed by six sample student responses, each in turn followed by the third type of rating prompt—"At which level in the *Linear Functions LP* would you place sample response [n]?"—and a set of five boxes, the first of which was labeled "Insufficient evidence" and the last four of which were labeled with Levels 1–4. Note that "insufficient evidence" was a rating option when assigning a LP level to a student response but not when assigning a level to an item. Examples from the rating forms for both MS1R and MS2R are provided in Appendix A.

## Procedure

The advance reading materials were e-mailed to the panelists 8 days prior to the meeting; panelists were asked to review the materials, contact us with any questions (none were received), and bring the materials to the meeting.

The meeting took place on Saturday, November 16, 2013, from 9:30 a.m. to 12:00 p.m. It began with a background presentation in which we provided a basic definition of a LP, that is, a theory of how student understanding about a concept or topic develops, and outlined the main goals of the research: to evaluate the validity of the theory and determine the extent to which different tasks provide evidence that a student is at a particular level.

We discussed the three different types of ratings panelists would be providing with respect to the tasks and student responses. When assigning a level from the LP to an item, panelists were asked to consider both the stem and the prompts and were reminded to supply a rationale. When assigning a level from the LP to a response, panelists were asked to consider the response in addition to the item stem and prompts. When rating the computational complexity of an item, we specified that panelists should consider both the difficulty of the numbers involved and the number of steps required in the expected solution (where “expected” solution refers to the solution most likely to be produced by an eighth-grade student, in the judgment of a panelist). We clarified that computational complexity refers to the complexity of the item for an eighth-grade student, not as determined by a computer algorithm. Finally, panelists were asked to ignore typographical errors in assigning their ratings.

The presentation concluded with a description of the linear functions LP (Arieli-Attali, 2011; Arieli-Attali *et al.*, 2012), which served to review and further clarify the materials that had been sent in advance (see also Graf & Arieli-Attali, 2015).

Panelists brought their own laptops to the session; wireless access was provided so that they could connect to the CBAL review page and view the tasks in their browsers. Following the presentation, panelists reviewed the MS1R task on their computers; in addition to reviewing the items, they were encouraged to interact with the task as a student might, to experiment with entering responses, and so on. Once they had finished reviewing MS1R, they provided ratings on the items and sample student responses on the corresponding paper forms by checking the appropriate boxes (see “Materials”). The screen prints on the forms served as reminders of the items they had reviewed on the computer. The panelists then reviewed the MS2R task on their computers and provided ratings on the MS2R paper form.

## Results

Before we discuss the results, we briefly summarize our research questions: The focus is on agreement among teachers’ LP level ratings of items and agreement of teachers’ LP level ratings with LP levels assigned by the test developer. In addition, we inspect agreement among teachers’ LP level ratings of sample response and the agreement with LP levels assigned by a trained rater.

All 10 invited teachers were present at the panel meeting. Because the mathematics assessment consists of many more items than the ELA assessment, we do not show the frequency distributions for all the ratings. They can be found in Appendix C.

### ***Computational Complexity Ratings***

Table 14 shows the statistics for the agreement among teachers on the computational complexity ratings for the 23 mathematics items. The same pattern arises as in the text difficulty ratings: Conger’s and Fleiss’s generalized kappas and Krippendorff’s alpha show similar (lower) values, and the Brennan – Prediger coefficient and Gwet’s AC2 show similar (higher) values. Using the interpretations in Table 3, the agreement among the teachers on the computational complexity of the mathematics items is judged as slight to moderate, which is slightly higher than the agreement on text difficulty found in the ELA study.

### ***Learning Progression Level Ratings for Mathematics Items***

In Table 15, the results for the agreement among teachers are shown for the LP level ratings of the mathematics items. The results are a little less spread out than those for ELA, and the standard errors are somewhat smaller as well. The agreement among mathematics teachers can be judged as fair to substantial.

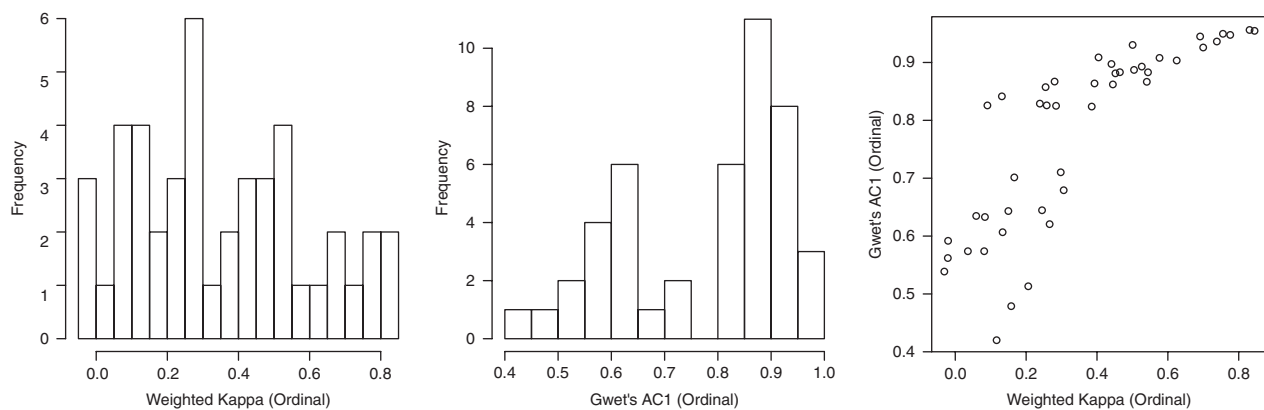
Because the results are mixed (although less so than for ELA), we show again the distributions of weighted kappa and Gwet’s AC1 for all 45 teacher pairs. The marginal and joint distributions are shown in Figure 2. As for ELA, it holds that weighted kappa is low if Gwet’s AC1, but not vice versa. Conversely, if weighted kappa is high, Gwet’s AC1 is high, but again, not vice versa. However, the scatterplot is less extreme than that for ELA, which is probably because the mathematics

**Table 14** Agreement Among Mathematics Teachers on Computational Complexity of Mathematics Items

Statistic	% Agreement		Estimate (SE)	95% CI		Agreement
	Observed	Chance				
Conger's generalized kappa	.83	.79	.21 (.06)	.08	.33	Fair
Brennan–Prediger coefficient	.83	.72	.41 (.05)	.31	.51	Moderate
Fleiss's generalized kappa	.83	.80	.18 (.07)	.04	.32	Slight
Krippendorff's alpha	.84	.80	.18 (.07)	.04	.33	Slight
Gwet's AC2	.83	.68	.48 (.05)	.38	.59	Moderate

**Table 15** Agreement Among Mathematics Teachers on Learning Progression Levels of Mathematics Items

Statistic	% Agreement		Estimate (SE)	95% CI		Agreement
	Observed	Chance				
Conger's generalized kappa	.86	.79	.31 (.05)	.20	.41	Fair
Brennan–Prediger coefficient	.86	.69	.55 (.04)	.46	.63	Moderate
Fleiss's generalized kappa	.86	.80	.29 (.05)	.18	.41	Fair
Krippendorff's alpha	.86	.80	.30 (.06)	.18	.41	Fair
Gwet's AC2	.86	.62	.62 (.06)	.51	.74	Substantial



**Figure 2** Distributions of agreement statistics for 45 teacher pairs for learning profession level ratings of mathematics items: (left) weighted kappa, (middle) Gwet's AC1, and (right) scatterplot.

teachers rated many more items than the ELA teachers (23 vs. 6 items). Note that Gwet's AC1 is never smaller than .40, indicating at least moderate agreement.

Table 16 displays the joint and marginal distributions of the LP level ratings of three different combinations of results: low weighted kappa–low Gwet's AC1, low weighted kappa–high Gwet's AC1, and high weighted kappa–high Gwet's AC1. For the first pair, the marginals are very different, so there is little possibility for high agreement. For the second, the marginals are not that different, but Cohen's kappa is still low (.13). For the third pair, 19 out of 23 ratings match exactly.

Table 17 shows the agreement on the LP level of the mathematics items of each teacher with the test developer's LP level. The agreement statistics are highest for Teacher 5, ranging from substantial to almost perfect, and lowest for Teacher 7, ranging from slight to substantial. Note that the individual agreements are substantially larger than found in the ELA study (see Table 9).

Table 18 shows the agreement results for the modal LP level teacher ratings and the LP level of the test developer for the mathematics items. Note that ordinal weights were used for all agreement statistics. The agreement is substantial to almost perfect, which is much higher than the agreement for ELA (see Table 10).

The results on the relations between computational complexity and LP level ratings of the items are shown in Appendix C.

**Table 16** Distributions of Three Different Pairs of Ratings of Mathematics Items Tasks That Lead to Different Agreement Statistics

Pair	LP level	LP level				Sum
		1	2	3	4	
Low weighted kappa – low Gwet’s AC1	1	0	0	0	0	0
	2	1	0	0	0	1
	3	8	8	0	0	16
	4	1	2	2	1	6
	Sum	10	10	2	1	23
Low weighted kappa – high Gwet’s AC1	1	0	0	1	0	1
	2	0	3	4	1	8
	3	0	4	6	1	11
	4	0	0	1	1	2
	Sum	0	7	12	3	22 <sup>a</sup>
High weighted kappa – high Gwet’s AC1	1	1	0	0	0	1
	2	0	7	2	0	9
	3	0	1	7	1	9
	4	0	0	0	4	4
	Sum	1	8	9	5	23

Note. LP = learning progression.

<sup>a</sup>One rating in this pair is missing.

**Table 17** Agreement Statistics for Individual Mathematics Teachers With Learning Progression Level of Test Developer

Teacher	Cohen’s kappa (SE)	Scott’s pi (SE)	Brennan – Prediger coefficient (SE)	Krippendorff’s alpha (SE)	Gwet’s AC1 (SE)
1	.48 (.12)	.45 (.14)	.80 (.06)	.47 (.14)	.86 (.04)
2	.68 (.10)	.68 (.10)	.88 (.04)	.69 (.10)	.92 (.03)
3	.60 (.15)	.59 (.16)	.83 (.06)	.60 (.16)	.88 (.04)
4	.19 (.19)	.17 (.20)	.71 (.08)	.19 (.20)	.82 (.07)
5	.72 (.12)	.72 (.12)	.90 (.04)	.72 (.12)	.93 (.02)
6	.67 (.15)	.66 (.15)	.86 (.06)	.67 (.15)	.90 (.04)
7	.21 (.13)	.05 (.22)	.52 (.10)	.07 (.22)	.63 (.08)
8	.21 (.14)	.18 (.17)	.56 (.11)	.20 (.17)	.65 (.10)
9	.24 (.14)	.20 (.17)	.68 (.08)	.22 (.17)	.78 (.06)
10	.33 (.13)	.25 (.18)	.74 (.08)	.27 (.18)	.85 (.06)

**Table 18** Agreement Statistics of Modal Learning Progression Level Rating of Mathematics Teachers With Learning Progression Level of Test Developer for Mathematics Items

Statistic	% Agreement		Estimate (SE)	95% CI	Agreement
	Observed	Chance			
Cohen’s kappa	.93	.80	.65 (.14)	.36 .93	Substantial
Scott’s pi	.93	.80	.64 (.15)	.33 .94	Substantial
Brennan – Prediger coefficient	.93	.69	.77 (.08)	.59 .95	Substantial
Krippendorff’s alpha	.93	.80	.65 (.15)	.34 .95	Substantial
Gwet’s AC1	.93	.62	.81 (.08)	.65 .97	Almost perfect

### Learning Progression Level Ratings for Mathematics Sample Responses

Table 19 shows the results for the agreement among mathematics teachers on the LP levels of the sample responses to the four mathematics items. The agreement is fair to moderate and considerably lower than the agreement among ELA teachers, which was substantial (see Table 12).

Table 20 shows the agreement results of the modal LP level rating of mathematics teachers with the LP level of the trained rater for six sample responses to each of four mathematics items. The agreement is moderate to substantial, which



**Table 19** Agreement Among Mathematics Teachers on Learning Progression Levels of Sample Responses to Mathematics Items

Statistic	% Agreement		Estimate ( <i>SE</i> )	95% CI		Agreement
	Observed	Chance				
Conger's generalized kappa	.86	.80	.33 (.07)	.17	.48	Fair
Brennan – Prediger coefficient	.86	.72	.51 (.05)	.40	.62	Moderate
Fleiss's generalized kappa	.86	.80	.31 (.08)	.14	.47	Fair
Krippendorff's alpha	.86	.80	.31 (.08)	.15	.48	Fair
Gwet's AC2	.86	.68	.57 (.06)	.45	.70	Moderate

**Table 20** Agreement of Modal Learning Progression Level Rating of Mathematics Teachers With Learning Progression Level of Trained Rater for Sample Responses

Statistic	% Agreement		Estimate ( <i>SE</i> )	95% CI		Agreement
	Observed	Chance				
Cohen's kappa	.91	.78	.57 (.13)	.31	.84	Moderate
Scott's pi	.91	.79	.56 (.15)	.25	.86	Moderate
Brennan – Prediger coefficient	.91	.72	.67 (.11)	.44	.90	Substantial
Krippendorff's alpha	.91	.79	.56 (.15)	.26	.87	Moderate
Gwet's AC1	.91	.64	.74 (.09)	.55	.94	Substantial

is slightly lower than found for ELA (see Table 13). However, the lower results are mostly due to the last item, where five out of six scores from the trained rater were zero (this was done to illustrate different types of responses that still were scored as zero). If the sample responses for this item are left out, then Cohen's kappa improves to .78 ( $SE = .10$ ) and Gwet's AC1 improves to .89 ( $SE = .05$ ).

## Summary

The agreement among the teachers on the computational complexity of the mathematics items was fair to moderate. On average, a substantial correlation was found between the teachers' ratings of the LP level and computational complexity of the items.

With respect to our first research question, the agreement among the teachers in assigning the levels to the items was fair to substantial. For the second research question, the agreement between the modal LP level of the teachers and that of the test developer was substantial to almost perfect.

With respect to the third research question, agreement among teachers of the LP levels of sample responses was fair to moderate. For the fourth research question, agreement of the modal LP level rating of the teachers with the LP levels assigned to the responses by a trained rater was substantial to almost perfect for three out of four items.

## Discussion

Our goal was to investigate how teachers interpret LPs for the purpose of assessment. In this study, we reported on an investigation in which teachers were asked to assign levels in LPs to ELA tasks, mathematics items, and sample responses for both. As outlined in the introduction, we evaluated four types of agreement with respect to level assignments from an LP. It was an open question as to how high the agreement among teacher ratings on both items and student responses would be, and as discussed, this agreement was not always good. A possible reason for this is the short period of time allocated to training on the LPs. For the argumentation LP, all four types of agreement were mixed. As noted, however, the ratings of the ELA team of test developers aligned with the P+ values, while the mode of the teacher ratings did not. In this case, further training might entail more in-depth coverage of the argumentation LP and all of its strands and improved understanding of how to weigh the strands collectively. For the linear functions LP, the positive finding that the modal teacher rating showed substantial agreement with the test developer's/trained rater's ratings suggests that even greater agreement might be conferred by sharpening the distinctions between levels through further training. In particular, the

use of both benchmark and rangefinder exemplars of student performances might be beneficial. Benchmarks fall exactly at a particular level, whereas rangefinders are borderline cases (e.g., high 2's or low 3's).

It is not easy to come to a final conclusion because some of the results are quite mixed (see the summaries) and also because different agreement statistics gave quite different results. Although it was not our intention to compare the ELA and mathematics results, it is interesting to see that the agreement on the LP levels of ELA tasks was generally worse than the agreement on the LP levels of mathematics items. This result may be explained by the fact that, at first sight, the argumentation LP with its different strands is more elaborate and complex than the linear functions LP. An alternate explanation is that the finding may be related to the substantial difference in numbers of items. Interestingly, for the sample responses, the opposite was found: The agreement on the LP levels of ELA sample responses was generally better than the agreement on the LP levels of mathematics sample responses.

As noted, the purpose of this study was largely exploratory, and our samples were not large. However, we think that the different results for ELA and mathematics on LP level ratings of items indicate that if an LP is more complex, sufficient time should be spent on familiarizing teachers with it. Also, it was found that different marginal distributions of ratings generally led to low agreement. We think this issue can be fixed with more training on the task at hand (either rating LP levels of items or rating LP levels of responses). For example, for the rating of responses, we did not have time available to provide the teachers with feedback on their ratings (typically, professional standard-setting procedures comprise at least two rounds of ratings, with feedback between rounds).

The main limitations of this study are that the sample size is generally limited and that the panel meeting was too short to provide satisfactory training and feedback. Nevertheless, some of the results are positive, and all results provide useful information for follow-up research (e.g., for determining sample sizes, see Gwet, 2014, §5.5). For similar research in the future, for example, we recommend allowing more time for feedback and for making the raters more familiar with the LPs. The latter is especially relevant if the LP under scrutiny is more elaborate.

## Acknowledgments

We would like to thank Lauren Phelps, Darlene Rosero, Melanie Schine, and Lynn Zaback for their work on recruitment and panel preparation; Patti Baron for providing advice on organizing standard setting panels and eliciting ratings; Nathan Lederer for setting up task review pages; Steven Holtzman and Jeff Wright for setting up the data file and data entry; and the panelists who provided the ratings for this study.

## Notes

- 1 Unless stated otherwise, by test developers, we mean both people who actually designed items and tasks and people who conduct validation research on the assessments.
- 2 The R code can be found online at [http://www.agreestat.com/r\\_functions.html](http://www.agreestat.com/r_functions.html)
- 3 Gwet (2014, Chapter 6) suggested determining cumulative probabilities of interval membership and using the highest agreement level for which this probability is at least .95. This approach is rather conservative, because it provides a lower bound. In addition, a normal approximation is used, which has been researched for some, but not all, agreement statistics (see, e.g., Gwet, 2008).
- 4 We initially present the results in aggregated form but split them out if necessary.
- 5 The  $P^+$  is defined as the mean score divided by the maximum possible score.

## References

- Alonzo, A., & Gotwals, A. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions*. Rotterdam, Netherlands: Sense. <https://doi.org/10.1007/978-94-6091-824-7>
- Altman, D. G. (1990). *Practical statistics for medical research*. New York, NY: CRC Press.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Arieli-Attali, M. (2011, June). *Middle school developmental models*. Presentation given to the IES Advisory Panel.
- Arieli-Attali, M., Wylie, E. C., & Bauer, M. I. (2012, April). *The use of three learning progressions in supporting formative assessment in middle school mathematics*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

- Bennett, R. E. (2011). *CBAL: Results from piloting innovative K–12 assessments* (Research Report No. RR-11-23). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2011.tb02259.x>
- Berry, K. J., & Mielke, P. W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, 48, 921–933. <https://doi.org/10.1177/0013164488484007>
- Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, 9, 71–123.
- Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687–699. <https://doi.org/10.1177/001316448104100307>
- Briggs, D. C., & Peck, F. A. (2015). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research and Perspectives*, 13, 75–99.
- Brillinger, D. R. (2004). Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics*, 18, 163–182.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage. <https://doi.org/10.4135/9781412985918>
- Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, 23, 31–50.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Confrey, J., Maloney, A., Nguyen, K., Mojica, G., & Myers, M. (2009). *Equipartitioning/splitting as a foundation of rational number reasoning using learning trajectories*. Paper presented at the 33rd annual conference of the International Group for the Psychology of Mathematics Education, Thessaloniki, Greece.
- Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88, 322–328. <https://doi.org/10.1037/0033-2909.88.2.322>
- Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (Research Report No. RR-63). Philadelphia, PA: Consortium for Policy Research in Education.
- Deane, P., Sabatini, J., & O'Reilly, T. (2012). *The CBAL English language arts (ELA) competency model and provisional learning progressions*. Retrieved from <https://www.ets.org/cbal/ela/>
- Deane, P., & Song, Y. (2014). *The key practice, "Discuss and Debate Ideas": Conceptual framework, literature review, and provisional learning progressions for argumentation* (Research Report No. RR-15-33). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/ets2.12079>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–430. <https://doi.org/10.1037/h0031619>
- Fleiss, J. L., Levin, B., & Paik, M. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York, NY: John Wiley.
- Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching*, 49, 1181–1210.
- Furtak, E. M., & Heredia, S. C. (2014). Exploring the influence of learning progressions in two teachers communities. *Journal of Research in Science Teaching*, 51, 982–1020.
- Gans, L. P., & Robertson, C. A. (1981). Distributions of Goodman and Kruskal's gamma and Spearman's rho in  $2 \times 2$  tables for small and moderate sample sizes. *Journal of the American Statistical Association*, 76, 942–946.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 732–764.
- Graf, E. A., & Arieli-Attali, M. (2015). Designing and developing assessments of complex thinking in mathematics for the middle grades. *Theory Into Practice*, 54, 195–202. <https://doi.org/10.1080/00405841.2015.1044365>
- Graf, E. A., & van Rijn, P. W. (2016). Learning progressions as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 165–189). New York, NY: Taylor & Francis.
- Gwet, K. L. (2008). Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika*, 73, 407–430. <https://doi.org/10.1007/s11336-007-9054-8>
- Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics.
- Janson, H., & Olsson, U. (2004). A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement*, 64, 62–70. <https://doi.org/10.1177/0013164403260195>
- Karelitz, T. M., & Budescu, D. V. (2013). The effect of the raters' marginal distributions on their matched agreement: A rescaling framework for interpreting kappa. *Multivariate Behavioral Research*, 48, 923–952. <https://doi.org/10.1080/00273171.2013.830064>
- Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, 30, 411–433.

- Landis, J., & Koch, G. (1977). Measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. <https://doi.org/10.2307/2529310>
- Li, P. (2016). A note on the linearly and quadratically weighted kappa coefficients. *Psychometrika*, 81, 795–801. <https://doi.org/10.1007/s11336-016-9501-5>
- National Research Council. (2006). *Systems for state science assessment*. Washington, DC: National Academies Press.
- National Research Council. (2007). *Taking science to school: Learning and teaching science in Grades K–8*. Washington, DC: National Academies Press.
- Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321–325. <https://doi.org/10.1086/266577>
- Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85, 257–268.
- Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, 26, 114–145. <https://doi.org/10.2307/749205>
- Smith, C., Wiser, M., Anderson, C., Krajcik, J., & Coppola, B. (2007). *Implications of research on children's learning for assessment: Matter and atomic molecular theory*. Paper commissioned by the Committee on Test Design for K–12 Science Achievement, Center for Education, National Research Council.
- Song, Y., Deane, P., Graf, E. A., & van Rijn, P. W. (2013). *Using argumentation learning progressions to support teaching and assessments of English language arts* (R & D Connections No. 22). Princeton, NJ: Educational Testing Service.
- Steedle, J., & Shavelson, R. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, 46, 669–715. <https://doi.org/10.1002/tea.20308>
- Tannenbaum, R. J., Baron, P. A., & Kannan, P. (2015). *Alignment between innovative summative assessment prototypes and the Common Core State Standards: An exploratory investigation* (Research Memorandum No. RM-15-07). Princeton, NJ: Educational Testing Service.
- van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Psicología Educativa*, 20, 109–115. <https://doi.org/10.1016/j.pse.2014.11.004>
- van Rijn, P. W., & Yan-Koo, Y. (2016). *Statistical results from the 2013 CBAL English language arts multistate study: Parallel forms for argumentative writing* (Research Memorandum No. RM-16-15). Princeton, NJ: Educational Testing Service.
- West, P., Rutstein, D., Mislevy, R., Liu, J., Levy, R., DiCerbo, K., ... Behrens, J. (2012). A Bayesian network approach to modeling learning progressions. In A. Alonzo & A. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 257–292). Rotterdam, Netherlands: Sense. <https://doi.org/10.1007/978-94-6091-824-712>
- Wilmot, D., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning*, 13, 259–291.
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46, 716–730. <https://doi.org/10.1002/tea.20318>
- Zwack, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, 103, 374–378. <https://doi.org/10.1037/0033-2909.103.3.374>

### Appendix A. Prompt for Ratings of First Task of Ban Ads

Ban Ads – Task 2 – Item 1

**Dear Editor:**

Advertising aimed at children under 12 should be allowed for several reasons.

First, one family in my neighborhood sits down and watches television together almost every evening. The whole family learns a lot, which shows that advertising for children is always a good thing because it brings families together.

Second, children can't remember commercials anyway, so they can't be doing kids any harm.

Finally, the arguments against advertising aren't very effective. Think about it: the advertising industry spends billions of dollars a year on ads for children. They wouldn't spend all that money if the ads weren't doing some good. Let's not hurt children by stopping a good thing.

If anyone doesn't like children's ads, the advertisers should just try to make them more interesting. The ads are allowed to be shown on TV, so they shouldn't be banned.

**Directions:** Read the "Dear Editor" letter. Then write a note for your classmates in which you identify and explain problems in the letter's *reasoning* or use of evidence.

Be sure to focus on just the most serious problems.

Type your answer in the box below.

Question for Item 1 in Task 2:

Consider the *Reasons and Evidence* learning progression. At which level in this progression would you place this item?

Level 1	Level 2	Level 3	Level 4	Level 5

### Appendix B. Prompt for Ratings of First Item of Moving Sidewalks—One Rider

Moving Sidewalks – One Rider: Item 1

Time: 0.0 sec

Stop when timer reaches:  sec  
 Stop when rider reaches:  ft

Use the Moving Sidewalk simulation to complete the table.

Time (seconds)	Distance of Ann From the Gate (feet)
0	0
2.5	<input style="width: 50px;" type="text"/>
<input style="width: 50px;" type="text"/>	18

**Questions for Item 1:**

At which level in the *Linear Functions* learning progression would you place this item?

Level 1	Level 2	Level 3	Level 4

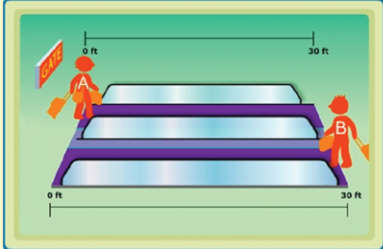
**Rationale:**

For an eighth grade student, what is the expected degree of *computational complexity* involved in responding to this item?

Not complex	Slightly complex	Somewhat complex	Moderately complex	Highly complex

Moving Sidewalks – Two Riders: Item 1

**CBAL MATH**



a. Why are the values in the Distance of Rider A column increasing as time increases?

b. Why are the values in the Distance of Rider B column decreasing as time increases?

Time (seconds)	Distance of Rider A from Gate (feet)	Distance of Rider B from Gate (feet)
4	8	22
6	12	18
10	20	10
13	26	4

**Questions for Item 1 (a-b):**

For the following questions, please consider parts 1a and 1b together as an item.

At which level in the *Linear Functions* learning progression would you place this item?

Level 1	Level 2	Level 3	Level 4

**Rationale:**

For an eighth grade student, what is the expected degree of *computational complexity* involved in responding to this item?

Not complex	Slightly complex	Somewhat complex	Moderately complex	Highly complex

### Appendix C. Additional Tables for Mathematics Study

Table C1 shows the frequency distribution of the teacher ratings of the computational complexity of the mathematics items.

**Table C1** Distributions of Teacher Ratings of Computational Complexity of Mathematics Items

Item	Computational complexity				
	Not	Slight	Somewhat	Moderate	High
MS1R_01	6	3	1	0	0
MS1R_02a	6	3	1	0	0
MS1R_02b	2	5	1	2	0
MS1R_03	7	1	1	1	0
MS1R_04ad	3	2	2	2	1
MS1R_05	2	3	2	1	2
MS1R_06	4	3	2	1	0
MS1R_07	1	5	4	0	0
MS1R_08ab	2	6	1	1	0
MS1R_8c	3	4	2	1	0
MS1R_10	1	1	2	5	1
MS2R_01ab	2	2	3	3	0
MS2R_02ab	1	2	2	5	0
MS2R_03ab	1	7	1	0	1
MS2R_04ab	3	4	1	2	0
MS2R_05	2	7	0	1	0
MS2R_06ab	1	6	2	1	0
MS2R_07ab	2	3	3	1	1
MS2R_08ab	2	4	1	2	1
MS2R_09	1	5	1	3	0
MS2R_10	1	1	5	3	0
MS2R_11ab	0	1	2	6	1
MS2R_12	0	0	3	5	2

Table C2 shows the student score frequencies for each of the 10 items in MS1R. These data are a subsample of data that were collected in an earlier study that focused on assessment of algebra for eighth- and ninth-grade students. They are used for reference here. The scores are directly linked to the LP levels (more details can be found in Graf & van Rijn, 2016). The second column indicates the LP levels that are used for scoring the student responses to the items. In Table C2, the distribution of the LP level ratings of the teachers is shown as well.

**Table C2** Student Score Frequencies and Teacher Rating Frequencies for Moving Sidewalks—One Rider

Item	Levels	Frequency				
		0	1	2	3	4
Student score <sup>a</sup>						
MS1R_01	1	35	165			
MS1R_02ab	1,2,3	45	58	71	26	
MS1R_03	2	46		154		
MS1R_04abcd	2,3	77		53	70	
MS1R_05	3	77			123	
MS1R_06ab	2,3	35		149	16	
MS1R_07ab	2	36		164		
MS1R_08abc	1	41	159			
MS1R_09ab	2	46		152		
MS1R_10ab	2,3,4	37		45	58	60



**Table C2** Continued

Item	Levels	Frequency				
		0	1	2	3	4
Teacher rating <sup>b</sup>						
MS1R_01			5	2	3	0
MS1R_02a			3	6	1	0
MS1R_02b			2	2	6	0
MS1R_03			2	3	5	0
MS1R_04abcd			2	1	7	0
MS1R_05			1	0	9	0
MS1R_06ab			1	5	4	0
MS1R_07ab			0	3	7	0
MS1R_08ab			0	7	3	0
MS1R_08c			0	3	6	0
MS1R_09ab			–	–	–	–
MS1R_10ab			1	0	0	9

<sup>a</sup>*n* = 200. <sup>b</sup>*n* = 10.

Table C3 shows the student score frequencies for each of the items in MS2R. This task was considerably more difficult than MS1R, as evidenced by the larger proportion of zero responses. In addition, the distribution of the LP level ratings of the teachers is shown. Note that there is one missing teacher rating for item MS2R\_4ab.

Table C4 shows the frequencies of the LP level ratings of the mathematics teachers for six sample responses to four mathematics items. The LP level ratings of the trained rater are shown as well.

**Table C3** Student Score Frequencies and Teacher Rating Frequencies for MS2R

Item	Levels	Frequency				
		0	1	2	3	4
Student score <sup>a</sup>						
MS2R_01ab	1,2,3	53	29	112	7	
MS2R_02ab	2	43		158		
MS2R_03ab	2	87		117		
MS2R_04ab	2,3	80		53	68	
MS2R_05	–	–	–	–	–	–
MS2R_06ab	2	86		115		
MS2R_07ab	2,3	91		56	54	
MS2R_08ab	3	151			50	
MS2R_09	2,3	66		70	65	
MS2R_10	2,3,4	78		36	54	33
MS2R_11ab	2,3	48		43	110	
MS2R_12ab	2,4	128		33		40
Teacher rating <sup>b</sup>						
MS2R_01ab			1	4	0	5
MS2R_02ab			0	1	4	5
MS2R_03ab			1	6	3	0
MS2R_04ab			1	0	8	0
MS2R_05			1	7	3	0
MS2R_06ab			0	6	4	0
MS2R_07ab			0	1	8	1
MS2R_08ab			0	4	5	1
MS2R_09			0	1	8	1
MS2R_10			0	1	2	7
MS2R_11ab			0	1	4	5
MS2R_12ab			0	2	2	6

<sup>a</sup>*n* = 201. <sup>b</sup>*n* = 10.

Table C5 shows Spearman’s rank order correlation, Goodman–Kruskal’s gamma correlation, and MI for the LP level and computational complexity ratings of the 10 teachers for the 11 items in MS1R. Although the values do give information about the relationship between the level and complexity ratings, they are quite unstable. For example, if we look at item MS1R\_10ab, we see that the rank order correlation is slightly negative, but this is caused by only one rater. So, even though we report these statistics, we are careful not to rely on too much interpretation of particular cases. However, there are substantial differences in the relations between the items: The ranges are  $-.19$  to  $.68$ ,  $-.60$  to  $1.00$ , and  $.05$  to  $.53$  for the

**Table C4** Frequencies of Learning Progression Level Ratings for Sample Responses to Four Mathematics Items

Item	Response	Trained rater LP level	LP level rating				
			0	1	2	3	4
MS1R_02a	1	2	1	3	5	1	0
	2	3	0	1	3	6	0
	3	1	3	3	4	0	0
	4	0	6	2	2	0	0
	5	3	1	0	2	7	0
	6	2	0	2	5	3	0
MS1R_05	1	2	1	0	6	3	0
	2	3	0	1	0	9	0
	3	0	3	3	3	1	0
	4	2	0	1	7	2	0
	5	3	0	0	0	9	1
	6	2	0	0	7	2	1
MS2R_01	1	2	1	0	6	0	3
	2	3	0	0	3	4	3
	3	0	8	2	0	0	0
	4	2	3	4	1	0	2
	5	3	1	2	2	2	3
	6	2	1	0	3	3	3
MS2R_08ab	1	0	1	3	3	3	0
	2	3	1	0	3	6	0
	3	0	1	1	7	1	0
	4	0	4	3	2	1	0
	5	0	3	1	5	1	0
	6	0	1	2	3	4	0

Note. LP = learning progression.

**Table C5** Spearman’s Rank Order Correlation ( $\rho$ ), Goodman–Kruskal’s Gamma Correlation ( $\gamma$ ), and Mutual Information for Relation Between Learning Progression Level Rating and Complexity Ratings for Moving Sidewalks—One Rider

Item	Max. LP level	$\hat{\rho}$ (SE)	$\hat{\gamma}$ (SE)	MI
MS1R_01	1	.31 (.21)	.43 (.40)	.46
MS1R_02a	3	.68 (.21)	1.00 (.00)	.48
MS1R_02b	3	.19 (.21)	.24 (.47)	.48
MS1R_03	2	.61 (.21)	1.00 (.00)	.27
MS1R_04abcd	3	.13 (.21)	.20 (.57)	.47
MS1R_05	3	.47 (.21)	1.00 (.00)	.19
MS1R_06ab	3	.60 (.21)	.75 (.26)	.53
MS1R_07	2	.00 (.21)	.00 (.58)	.05
MS1R_08ab	1	.39 (.21)	.69 (.33)	.23
MS1R_08c	1	.30 (.21)	.44 (.43)	.12
MS1R_09ab	2	–	–	–
MS1R_10ab	4	–.19 (.21)	–.60 (.36)	.07

Note.  $n = 10$ . LP = learning progression. MI = mutual information.

**Table C6** Spearman's Rank Order Correlation ( $\rho$ ), Goodman–Kruskal's Gamma Correlation ( $\gamma$ ), and Mutual Information for Relation Between Learning Progression Level Rating and Complexity Ratings for Moving Sidewalks—Two Riders

Item	Max. LP level	$\hat{\rho}$ (SE)	$\hat{\gamma}$ (SE)	MI
MS2R_01ab	3	.85 (.22)	1.00 (.00)	.61
MS2R_02ab	2	-.12 (.22)	-.14 (.48)	.33
MS2R_03ab	2	-.08 (.22)	-.06 (.61)	.48
MS2R_04ab	3	.44 (.22)	1.00 (.00)	.14
MS2R_05	–	.38 (.22)	.66 (.36)	.24
MS2R_06ab	2	.44 (.22)	.73 (.31)	.15
MS2R_07ab	3	.48 (.22)	.73 (.27)	.45
MS2R_08ab	3	.47 (.22)	.58 (.32)	.53
MS2R_09	3	.42 (.22)	.82 (.21)	.20
MS2R_10	4	-.17 (.22)	-.29 (.43)	.14
MS2R_11ab	3	.75 (.22)	1.00 (.00)	.56
MS2R_12	4	.81 (.22)	1.00 (.00)	.51

*Note.*  $n = 10$ . LP = learning progression. MI = mutual information.

rank order correlation, gamma correlation, and MI, respectively. Also, different values for the correlations can go together with nearly equal values for the MI statistics (see, e.g., MS1R\_02a and MS1R\_02b).

Table C6 shows Spearman's rank order correlation, Goodman–Kruskal's gamma correlation, and MI for the LP level and computational complexity ratings of the 10 teachers for the 12 items in MS2R. The averages of these statistics are slightly higher than the averages for MS1R.

### Suggested citation

van Rijn, P., Graf, E. A., Arieli-Attali, M., & Song, Y. (2018). *Agreement of teachers on evaluating assessments of learning progressions in English language arts and mathematics* (Research Report No. RR-18-11). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/ets2.12199>

**Action Editor:** Heather Buzick

**Reviewers:** Lili Yao and Randy Bennett

CBAL, ETS, the ETS logo, and MEASURING THE POWER OF LEARNING are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at <http://search.ets.org/researcher/>