Evaluating Learning Progression-Based Mathematics Rubrics for Validity and Reliability

Edith Aurora Graf, ETS, agraf@ets.org
Cheryl L. Lizano, Southern Illinois University Edwardsville, ceames@siue.edu
Francis Agyapong, Southern Illinois University Edwardsville, fagyapo@siue.edu
Peter W. van Rijn, ETS Global, pvanrijn@etsglobal.org

Abstract: We describe a validity and reliability examination of learning progression-based mathematics rubrics. Although rubric agreements were generally high, this does not guarantee a valid rubric. Nevertheless, lower agreements are indicative of possible validity and reliability issues, and we trace through a rubric with relatively lower agreement. We concluded that the rubric should provide more example responses. While agreements are useful indicators it is only through the process of reflection and discussion that validity issues with rubrics are uncovered.

Background and purpose

A learning progression (LP) is a domain-specific theory of how student thinking develops, from early ideas to the target understanding (e.g., Corcoran et al., 2009; Daro et al., 2011). An LP can serve as a guide for formative assessment design (Confrey, 2019; Confrey et al., 2020; Ketterlin-Geller et al., 2020). LPs are expressed as a sequence of levels, which are assumed to be ordered and distinct. These assumptions need to be empirically validated (Confrey, 2020; Graf & van Rijn, 2016; Wilmot et al., 2011). In our approach, responses are mapped to levels; this requires not only an LP, but task-specific rubrics: the rubrics are aligned with the LP and indicate how to map each response to each question to a level. In this paper we examine these rubrics for validity and reliability.

In our work we focus on students and schools that have been historically underserved in mathematics education. In the design and validation of an LP, it is important to include the perspectives of students the LP is intended to serve. We shared draft tasks with students during cognitive interviews and focus groups and revised them in accordance with what we learned (Graf et al., 2021). These tasks were designed for administration to individuals, however. As part of the Algebra Project's *Five-step curricular process* (Moses et al., 1989), students co-create mathematical understanding while working in teams. We therefore adapted the tasks for collaboration.

The concept of function learning progression

We focus on two strands of the concept of function LP: The Traditional strand, which focuses on *formula-based* functions (Carraher et al., 2008), and the Finite-to-Finite strand, which focuses on mappings from finite sets to finite sets and is aligned with the Algebra Project's *Road Coloring* curriculum module (Budzban & Moses, 2017). The research basis for the LP, which includes APOS Theory (Dubinsky & Wilson, 2013) is provided in Eames et al., (2021) and Graf et al. (2021). We present only an abbreviated version here, which summarizes the Traditional and Finite-to-Finite strands:

- Level 1: Students have an intuitive grasp of one-to-one functions; they can extend sequences and patterns as well as evaluate simple expressions.
- Level 2: Students think of a function as a formula; they focus on points rather than trends. They may not accept functions that do not show a pattern as such.
- Level 3: Students think of a function as a rule, which may or may not be expressed as a formula. They can translate between directed graphs and arrow diagrams.
- Level 4: Students have an appreciation of the uniqueness property of a function.
- Level 5: Students understand domain and range; compose and invert functions with understanding; work with function families; and translate from tables, graphs, or descriptions to equations.

Methodology

Participants

In total, 64 9th-grade students in six classes from a single school participated. The students were distributed across 20 teams. We did not collect demographic data for the sample, however publicly available data from NCES (nces.ed.gov) indicates that more than 95% of students at the school identify as either Black or Hispanic and more than 75% qualify for free or reduced-price lunch.

Instrumentation

Three parallel pairs of tasks were developed. Each task consists of five to six questions. The pairs include *Backpacks* and *Book Orders*, *Jogging Paths* and *Trail Map*, and *Restaurant Tables* and *Shape Patterns*. All tasks were designed based on an LP for the concept of function. Within each pair, the tasks assess identical concepts and skills. Rubrics for each task were also developed. For each response to each question, the rubrics indicate the level from the LP the response should receive.

Design and procedure

The study took place across four 40-minute periods on consecutive days. Data were collected by computer on Days 1, 2, and 4. Each class was assigned a pair of tasks. On Day 1, students responded to a task individually. On Day 2, students responded to the same task while engaging in three to four person teams in an online chat. On Day 3, the teacher led a class discussion of the same task. On Day 4, students responded to the parallel task individually. Each pair of tasks was seen by two classes. The order of the tasks was counterbalanced so that one class saw one task from the pair on Days 1, 2, and 3 and the other on Day 4; for the other class this was reversed.

Coding and analysis

Each task consists of several questions in a variety of response types, some of which are suitable for automated scoring. The remainder of the questions were scored independently by two pairs of raters. In this paper we focus on the human-scored questions. For each human-scored question, each pair of raters assigned a level of the LP and made notes about whether the rubric was clear and comprehensive.

LP scores were assigned to responses from both individual phases (Days 1 and 4) and the team phase (Day 2). We pooled the LP scores from both individuals and teams. Then, for each question, percent agreement and quadratically-weighted kappa were calculated. Low agreements can indicate that a rubric is unclear or invalid. However, high agreements do not indicate that a rubric is valid (Moskal & Leydens, 2019): We had notes concerning the validity of rubrics for questions where the agreement was high.

Results and discussion

Table 1 shows the rater agreements for the twelve human-scored items from the six tasks. Both the percent agreement and quadratically-weighted kappa are shown. Although the sample sizes are small, the kappas are generally good, with all but two values above .80.

Challenges with the rubrics resulted in two types of revisions. First, both teams noted incorrect responses that suggested partial understanding not acknowledged by the rubrics. So, these rubrics were revised to expand the space of predicted responses. Second, both teams noted that the mapping of some types of predicted responses to LP levels was inconsistent across items, so some rubrics were revised by shifting LP level assignments. In

Table 1, items requiring one of these revisions are indicated as minor revision, and items requiring two are indicated as major revision.

Table 1
Agreement Statistics

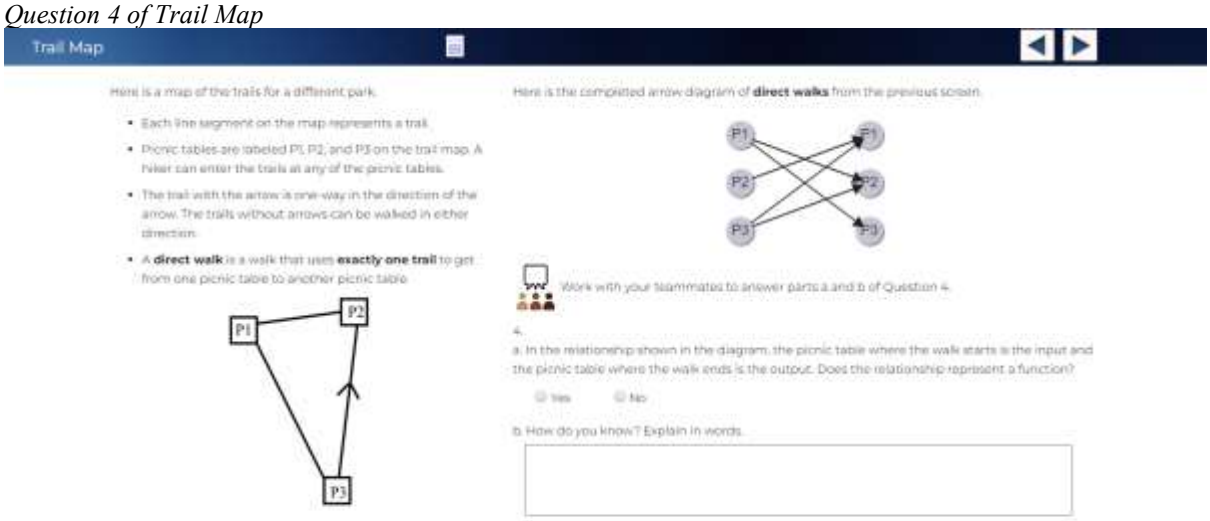
Task	Question	*N*	LP Levels	% Agreement	Weighted Kappa	Revisions
Backpacks	2	26	2	92.3	0.82	No revision
Backpacks	4	25	3	92.0	0.96	Major revision
Backpacks	5	25	3	92.0	0.94	Major revision
Book Orders	2	19	2	100.0	1.00	No revision
Book Orders	4	19	3	89.5	0.94	Major revision
Book Orders	5	18	3	72.2	0.83	Major revision
Jogging Paths	4	22	3	77.3	0.86	Minor revision
Restaurant Tables	2	20	4	80.0	0.90	Minor revision
Restaurant Tables	3	20	4	75.0	0.72	Major revision

Shape Patterns	2	18	4	88.9	0.84	Minor revision
Shape Patterns	3	18	3	83.3	0.80	Major revision
Trail Map	4	25	3	68.0	0.72	Minor revision

Next, we consider a question with relatively lower agreement, *Trail Map* Question 4 (Figure 1). Note that in the previous question, students were asked to create an arrow diagram showing all possible direct walks. Question 3 and Question 4 of *Jogging Paths* were similar to Question 3 and Question 4 of *Trail Map*, except that the scenario involved water stations connected by jogging paths and a slightly different map was used.

Figure 1

Question 4 of Trail Map



We noticed that many responses were decontextualized; students did not necessarily mention picnic tables, water stations, or arrows. For example, for Question 4 of Trail Map, one team wrote "It's a function because wherever a person starts will determine where they ended up, an input-output relationship is a function." Responses like this were difficult to interpret using the rubric, leading to low agreement. Adding decontextualized examples to the rubric for Question 4 of Jogging Paths and Trail Map could improve both validity and reliability. A revised rubric is shown in Table 2 (additions are shown in italics)**Error! Reference source not found.**.

Table 2

Revised Rubric for Trail Map Question 4

Level	Part a	Part b
Level 4	No	There are two arrows leaving P1; There are two arrows leaving P3;There are two arrows leaving P1 and P3; *Two inputs have more than one output; One input can't have multiple outputs*
Level 3	No	There are two arrows going to P2; There are two arrows going to P1; There are two arrows going to P2 and P1; *There are two arrows going to an output; You should only have 1 output and 1 input for each of the 3 picnic tables*
Level 3	Yes	You can get to every picnic table/There are arrows going to each picnic table; You can get to any other picnic table from any starting point; *Wherever a person starts will determine where they ended up (if the arrow diagram from Question 3 shows a function)*
Not yet at Level 3	No	You can't get directly to P3 from P2 without going through P1; *An output can't have multiple inputs, but an input can have multiple outputs;* Any other response
Not yet at Level 3	Yes	All of the picnic tables are used up; *You have an output and also an input;* Any other response

Conclusions and implications

Before empirically validating an LP, the rubrics must be evaluated. We noted validity issues in the rubrics while scoring, and we examined agreements as indicators of insufficient validity and/or reliability. We noted validity issues even in tasks with high agreement. While the rubrics for *Trail Map* and *Jogging Trails* were highly specific, the responses were often decontextualized. The rubrics as written were insufficient to capture these abstractions and revised accordingly. While agreements are useful indicators of potential issues, it is only through the process of reflection and discussion that the nature of these issues are uncovered.

References

Budzban, G., & Moses, R. (2017). Road coloring. Retrieved from https://iris.siue.edu/math-literacy-archive/items/show/307

Carraher, D. W., Martinez, M. V., & Schliemann, A. D. (2008). Early algebra and mathematical generalization. *ZDM Mathematics Education, 40*(1), 3–22. https://doi.org/10.1007/s11858-007-0067-7

Confrey, J. (2019). A synthesis of research on learning trajectories/progressions in mathematics. Retrieved from Organization for Economic Co-operation and Development https://www.oecd.org/education/2030-project/about/documents/A_Synthesis_of_Research_on_Learning_Trajectories_Progressions_in_Mathematics.pdf

Confrey, J., Toutkoushian, E., Shah, M. (2020). Working at scale to initiate ongoing validation of learning trajectory-based classroom assessments for middle grade mathematics. *Journal of Mathematical Behavior*, *60*, 100818. http://dx.doi.org/10.1016/j.jmathb.2020.100818

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). Learning progressions in science: An evidence-based approach to reform (Research Report #RR-63). Philadelphia: Consortium for Policy Research in Education. https://doi.org/10.12698/cpre.2009.rr63

Daro, P., Mosher, F. A., & Corcoran, T. B. (2011). Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction. *CPRE Research Reports*. Retrieved from http://repository.upenn.edu/cpre_researchreports/60

Dubinsky, E., & Wilson, R. T. (2013). High school students' understanding of the function concept. *The Journal of Mathematical Behavior, 32*, 83–101. https://doi.org/10.1016/j.jmathb.2012.12.001

Eames, C. L., Graf, E. A., van Rijn, P. W., Budzban, G., & Voepel, T. (2021). The finite-to-finite strand of a learning progression for the concept of function: A research synthesis and cognitive analysis. *The Journal of Mathematical Behavior, 62,* 100864. https://doi.org/10.1016/j.jmathb.2021.100864

Graf, E. A., van Rijn, P. W., & Eames, C. L. (2021). A cycle for validating a learning progression illustrated with an example from the concept of function. *The Journal of Mathematical Behavior, 62,* 100836. https://doi.org/10.1016/j.jmathb.2020.100836

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: Taylor & Francis.

Ketterlin-Geller, L. R., Zannou, Y., Sparks, A., & Perry, L. (2020). Empirical recovery of learning progressions through the lens of educators. *The Journal of Mathematical Behavior, 60*, 100805. http://dx.doi.org/10.1016/j.jmathb.2020.100805

Moses, R., Kamii, M., Swap, S. M., & Howard, J. (1989). The algebra project: Organizing in the spirit of Ella. *Harvard Educational Review*, *59*(4), 423–444. https://doi.org/10.17763/haer.59.4.27402485mqv20582

Moskal, B. M. and Leydens, J. A. (2019). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation*, *7*(10). https://doi.org/10.7275/q7rm-gg74

Wilmot, D. B., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning, 13*(4), 259–291. https://doi.org/10.1080/10986065.2011.608344

Acknowledgments

The research reported here was supported by the National Science Foundation, through grant 2101393 to Educational Testing Service. The opinions expressed are those of the authors and do not represent the views of

the National Science Foundation. We wish to extend thanks to all staff at ETS, the Algebra Project, YPP, SIUE, and UNL who contributed to this work, and to staff and students at the participating school.