https://doi.org/10.36315/2025v1end095
Education and New Developments 2025

TEAR-FREE MATH ASSESSMENT FOR INFORMAL LEARNING RESEARCH

Teresa Lara-Meloy, Ken Rafanan, & Elise Levin-Güracar
TERC, Inc. (USA)

Abstract

Assessments are often necessary for justifying making-based education to stakeholders, but traditional assessments rarely align with the values and learning goals of out-of-school time (OST) making-based learning programs. In fact, assessments require that the youth being evaluated genuinely participate in the assessment, something that might be rare or difficult in OST making programs. In this paper, we describe effective, culturally-responsive mathematics assessments designed for youth in OST maker programs and the rationale that led to the assessment development. We then share the design framework that we used to develop these assessments to be camouflaged, appearing to learners to be like other making tasks. We share the results from our pilot with middle schoolers during an afterschool implementation of a math+making program in California, USA. The assessment data relate to measurement, volume, tessellations and transformations, and symmetry. Our research shows that more open forms of hands-on assessments can provide each and every student the opportunity to show what they have learned and leave with something they have made.

Keywords: *Mathematics, assessment, maker education, cultural responsivity, out-of-school learning.*

1. Introduction

Afterschool time offers a tantalizing opportunity to engage youth in rich, meaningful, and fun STEM activities, with unstructured time and few constraints on the type of activities offered. Afterschool is the perfect setting for an ambitious making-mathematics-community program, like Amped 4 Making (A4M). In the four semesters, middle schoolers use mathematics to design, prototype, and build meaningful and useful artifacts. Throughout they learn about and embody principles derived from the life of Cesar Chavez, such as service-orientation, appreciation of their communities, and development of persistence (also a math practice!). The question for such programs as ours is whether youth indeed are learning math or other STEM content knowledge.

2. Background/Lit

In the early 2000s, the National Science Foundation (NSF) made sizeable investments into developing, supporting and researching models of STEM learning in out-of-school-time (OST) learning environments. Practitioners and researchers soon recognized the challenges of assessing learning in these programs. In 2009, NSF sponsored an OST STEM conference with over a hundred participants from the field, and a major finding from the research and evaluation strand was “the need to develop embedded and performance assessments that could document students’ development of skills or understanding in ways that did not interfere with students’ experience of the OST STEM programs” (Bevan et al., 2013, p. 15). Numerous efforts were undertaken to research and develop invisible or unobtrusive assessments. Pattison, Rubin, and Wright (2016) discussed how the context of the assessment matter. Fu, Kannan, and Shavelson (2019) discuss the goal of assessment while minimizing any disruption to the OST experience. The assessment tasks presented in this paper align with the work of others in the field to fully center participants’ learnings while also assessing their development of mathematics content.

3. Design

One of our research questions was "What math content and practices do youth learn through participation in AMPD4Math?" To answer this question, we proposed and piloted an existing instrument to assess mathematical content knowledge, the MARS Tasks (MAP, 2007), presupposing that it was appropriate to the informal education setting because of its participatory nature. The MARS Tasks are commonly used as formative assessments in classrooms and are known to reliably show changes in students' math content knowledge and reasoning skills. Our team of researchers and designers reviewed the publicly available set of MARS Tasks on the MAP website and selected a subset that aligned with the mathematics content in the first semester of A4M. The first semester of A4M culminated in a group project where each participant built a cubby for storing maker supplies and projects. We chose a task assessing volume and surface area, Smoothie Box, because it targeted the lowest of the grades we served and would be possibly accessible to all our youth. We administered the task as pre and post assessments in our pilot year. While 16 students took the pre-assessment across both sites, only 5 students took the post-assessment and only at one site. Facilitators at the site with no post-assessment participants reported that because the students actively disliked taking it, the site staff decided not to administer it again. Of the 5 students who took a post-assessment, only 2 had a taken pre-test. Thus, no statistical analysis on pre-post growth was possible.

Importantly, while there is a socio-institutional contract during school hours between teachers and students regarding how we assess learning through testing, no such agreement would hold in informal learning settings such as our afterschool programs. There was no external incentive in A4M to encourage active participation in testing, such as grades. And, few youths showed the internal motivation to participate in what to them seemed like just another math test after a full day of school and during what is often considered fun and free times.

Without an appropriate existing assessment to address our research question, we took the feedback from youth and facilitators and developed new assessments using a mixture of evidence-centered approach (Mislevy, R. J., & Haertel, G. D., 2006) and a design approach (Goldman & Kabayadondo, 2016). Our design principles were that the assessment needed to:

(1) assess the mathematics relevant to the major work of an A4M semester.
(2) require hands-on making, so that the task doesn't feel like an assessment.
(3) result in a product/artifact that the learner can take with them.
(4) have a feedback mechanism within the task to enable learners to check their own work.
(5) be culturally relevant.
(6) take approximately 20 minutes to complete.

For each semester, we created a bank of ideas for the assessments and evaluated them against our design principles. For finalist items, we refined the language to exclude confusing terminology and provided definitions where needed. We tested the item first with other team members and then with small numbers of readily-accessible youth in the same age-range as the target youth. The result was what we call Maker Tasks.

Table 1. Maker Tasks related to Math Content and Curricular connection.

Maker Task	Math Content	Curriculum Tie-in
1. Box for Dice	Volume (discrete) of objects with linear measurements	Cubbie storage
2. Hojalata Repujada	Symmetry	Symmetrical designs on luminaries
3. Carved eraser stamp	Rigid transformations	Paper circuit designs
4. Bean container	Volume (continuous)	Filling in the raised planter bed with soil

4. Methods

The Maker Tasks were embedded in the semester-long programs. Youth completed a maker task two times each semester: once at the beginning and again at the end. Participants worked independently, but sat at group tables, with access to the necessary materials. Facilitators were given a script to encourage youth to work individually and try as best they could. Then facilitators read the maker task out loud, in English, and if they could, in Spanish. Facilitators were asked not to give hints and to redirect youth to the text of the task. Staff collected the artifacts from the tasks for scoring. In the case of the Dice box and the Bean container, facilitators collected all the boxes/containers. In the case of the Hojalata and Carved Eraser

Stamp, only the paper drafts/sheets were collected and youth were able to keep the aluminum sheet artifact and the eraser stamp.

Staff collected the Maker Task artifacts from the program sites and anonymized each artifact keeping a key list and then sent the artifacts to the research team for scoring. A team of three experienced math education researchers developed and refined the scoring rubrics with un-paired maker tasks. We set used a small set of artifacts to train, ensure inter-rater reliability and sharpen decision points in the rubric. In the end, because we had fewer than expected maker tasks to score, all three researchers scored all tasks. Any conflicts were resolved by consensus in conversations.

5. Maker task results

Figure 1. Example of a student solution to Maker Task 1: Box for Dice.

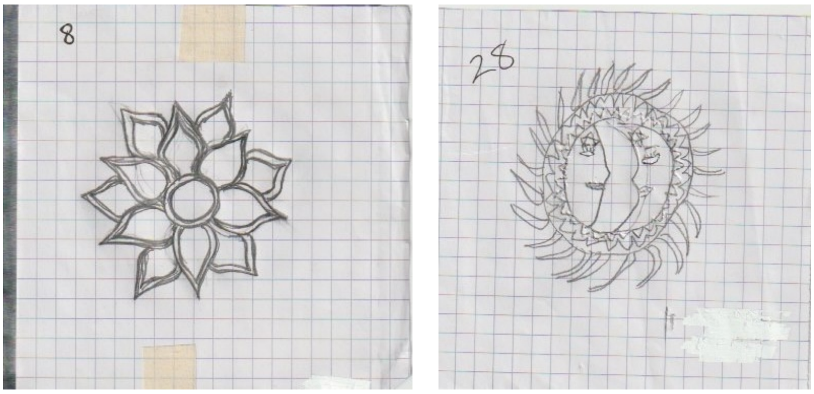
In Fall 2023, the Maker Task involved youth creating a box to contain a set of 12 dice at the beginning and at the end of the program. We scored the pre- and post- for the 19 youth who completed both a pre- and post-Maker Task. The tasks were randomly numbered so they could be scored without researchers knowing if they were a pre- or post-assessment. They were then scored by at least two researchers to ensure consistency in the scoring. Researchers used a rubric that was aligned with the task directions and embedded math and making concepts.

Table 2. Rubric to Score Maker Task 1: Box for Dice.

Is it a container?	(0, 1)
Does it have a top?	(0,1)
Does it hold 12 dice so they do not shake around?	(0, 1)
Does it not have overlapping sides?	0 (it has overlapping sides) 1 (it doesn't)
Does it have the correct height, width, and length of the sides marked on the outside of the box?	0 (no) 1 (yes, correct) 1 for each
Precision of construction	0 (not remotely a box), 1, 2

Math and making are entwined in the task, so math cannot be disambiguated from the making. Overall growth across all rubric items was statistically significant from pre- to post- ($p < .05$). While 79% of youth made a container with a top in the pre-assessment (which was one of the assessment's requirements), 100% of youth made a container with a top by the post. Additionally, the construction improved vastly. In the pre-assessment, 37% of youth created a container where the dice didn't shake around (which was another requirement), while 84% of youth were able to by the post-assessment. Youth also improved in having correct measurements (length, width, and height) marked on their box (another requirement). On the pre-assessment only 26% of youth had at least one correct measurement and only one youth had all three correct. On the post-assessment 73% had at least one correct measurement, and three youth had all three correct.

Figure 2. Examples of student Hojalata prototypes for Maker Task 2: Hojalata Repujada.



In Spring 2024, youth were asked to make an Hojalata Repujada, a metal embossed design, with at least one line of symmetry. Overall, we scored the 21 tasks for youth with both pre- and post-Maker Tasks, using the same protocol as in semester 1. We created a unique scoring rubric for this task. In the pre-, 43% youth had no symmetry in their design. This dropped to 10% in the post. There was also a decrease in average complexity from pre to post (-.38), which may have been because youth were more aware of trying to make designs that they would be able to add symmetry to successfully. We examined partial scores (symmetry and symmetry precision) since this was the math content we were interested in and found the growth from pre to post to be statistically significant ($p < .05$) on a 2-tailed t-test. Changes in overall scores (including complexity) were not statistically significant at a $p < .05$ level.

Table 3. Rubric to Score the Hojalata Repujada Maker Task.

Symmetry	(No=0, Some=1, Very=2)
Symmetry Precision	(Roughly=1, Moderate=2, Very=3, No Symmetry=NA)
Complexity	(Simple=1, Moderate=2, Complex=3)

While we have a small total N across the Fall 2023 and Spring 2024 administrations, we consider that the clear growth from pre to post represents the utility of using the Maker Tasks again in Year 3, since the Maker Task have been found sensitive enough to reflect that growth.

6. Discussion

The team considers the development of Maker Tasks to be a productive direction for our research as our observations and assessment data analysis indicate that they meet our dual needs of assessing what mathematics participants can do while the tasks appear to those participants to be a making task that is in line with the other making that happens in the program. Indeed, in some end of semester interviews some participants mentioned the Maker Tasks to be some of their favorite activities. In program reflections at end of the semester, youth interviewed were able to verbalize the difference in their own making ability and explain how they had improved in their post-maker task. Our small sample sizes show growth in three out of the four tasks. Unfortunately, our project was terminated early by the US government shake up and we could not confirm this trend with our Container for Beans task.

Maker Tasks have proven a much more natural assessment activity for the program's context because it engages youth in making, which is the program's central activity. It is worth noting that this making aspect of the task can also raise challenges to scaled-up implementations. In making tasks, the exact materials can make a difference. Printer paper does not fold the same as cardstock, not all dice are the same size (even when specifications are provided on a seller's website), the gauge of aluminum for an Hojalata can require more or less pressure for embossing, etc. Although scoring the Maker Task artifacts generally requires about the same amount of effort as scoring a MARS task response, they do require more effort than a typical paper-and-pencil math test. Despite these extra challenges, for OST programs like A4M, they are an invaluable research tool to have available.

7. Conclusions

While more research is necessary, Maker Tasks offer a promising opportunity for assessing targeted STEM concepts and skills. However, they take time to develop and refine. Once in use, however, these tasks not only provide insight to researchers and stakeholders interested in knowing whether youth are developing STEM content knowledge but also offer tangible feedback and an enjoyable making experience to the participants themselves.

Acknowledgments

This material is partially based upon work supported by the National Science Foundation. AMPED 4 Making was funded through the AISL Program (award #2215382) until it was prematurely terminated. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
We are grateful to the middle schoolers who graciously attempted the first assessment and gave us their feedback as well as the middle schoolers who engaged with the Maker Tasks. With similar gratitude we recognize the work of our facilitators in administering, observing and reporting on youth engagement, and collecting the assessments.

References

Bevan, B., Michalchik, V., Bhanot, R., Rauch, N., Remold, J., Semper, R., & Shields, P. (2010*). Out-of-school time STEM: Building experience, building bridges*. San Francisco, CA: Exploratorium. Retrieved April, 29, 2013.

Fu, A. C., Kannan, A., & Shavelson, R. J. (2019). Direct and unobtrusive measures of informal STEM education outcomes. In A. C. Fu, A. Kannan, & R. J. Shavelson (Eds.), Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. *New Directions for Evaluation*, *161*, 35-57.

Goldman, S., & Kabayadondo, Z. (Eds.). (2016). *Taking design thinking to school: How the technology of design can transform teachers, learners, and classrooms*. Taylor & Francis.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational measurement: issues and practice, 25*(4), 6-20.

Pattison, S., Rubin, A., & Wright, T. (2016). *Mathematics in informal learning environments: A summary of the literature. Institute for Learning Innovation.* Math in the Making Project. TERC.