Request for Proposals #1 – Formative Assessment

Due Date: March 8, 2026 (Midnight Pacific Standard Time)
Application Link: https://dp.skipso.com/en/custom/k12aiapplication/new/submit

This is the first of several anticipated requests for proposals (RFPs) from the K-12 AI Infrastructure Program. Overall, the RFPs will fund teams to produce datasets, benchmarks, and/or models as public goods, intending to support improvements to multiple applications of artificial intelligence (AI) in K-12 education. This first RFP focuses on K-12 formative assessment and supports two tracks: **Track 1** for proof of concept projects and **Track 2** to enhance an existing asset.

Vision

Our vision in this RFP is to produce public goods that enable a wide variety of products and services to provide high-quality formative assessment to every student, with particular attention to the needs of U.S. students furthest from opportunity, thereby accelerating student success in multiple K-12 subject areas. The public goods will be licensed resources that can be broadly used and incorporated—directly or indirectly—by developers of AI-enabled tools and infrastructure for K-12 education, in order to improve the technology products that teachers and students rely on.

The logic model of our K-12 AI Program Team has four phases:

1. **Soliciting and funding the production of public goods** through this RFP and subsequent RFPs.
2. **Providing grantee support** during the period of performance to enable rapid progress towards high quality, low risk public goods and to forge fruitful relationships among the community of awardees.
3. **Fostering adoption and use** mostly after the period of performance (though some early adopter engagement may begin during funded projects) to attract intended users to awareness and adoption of newly-available public goods.
4. **Tracking uptake and impacts** as public goods become incorporated into a multitude of products and services, which can be via pre-trained datasets, benchmarks, Retrieval-Augmented Generation (RAG) systems, Model Context Protocol (MCP) servers, and other emergent methods that incorporate a public good into a specific product or service to improve its educational value.

This logic model reflects a theory of change in which well-designed public goods, combined with early technical support and downstream adoption efforts, lead to widespread improvements in AI-enabled formative assessment practice.

Multimodality

We see the future of AI-enhanced formative assessment as based in multimodal interactions with students and teachers. Including more modalities, such as speaking, listening, drawing, writing, etc., can enable students to better demonstrate what they know and can do, which can lead to formative assessments that are more accurate and fair. Multiple modalities can respond to learner variability of many kinds. "Multimodal" thus specifically includes spoken communication, sketches, diagrams and other visuals, and perhaps tangible interactions. We anticipate continued use of clickstream, keystroke, and textual data and look towards growing opportunities to integrate commonplace and emerging modalities to advance AI-enabled formative assessment. *Note: We are not presently investing in student or classroom video, or sensor-based approaches (e.g. eye-tracking) that face challenges of practicality or heightened privacy concerns.*

Principles

We further envision advancing AI-enhanced formative assessment by adhering to the following principles:

- **Grounding in the learning sciences** and operationalizing learning sciences constructs and techniques for (a) understanding what students know and can do and (b) designing adaptive plans for instructional next steps. There is a large, mature research literature on formative assessment and feedback research that should inform the design of technical solutions. We are investing in operationalizing research-based insights in a public good, not funding new basic or foundational learning sciences research.
- **Designing for Targeted Universalism** to serve every learner while intentionally attending to specific population(s) that would strongly benefit from formative assessment that is tuned to their strengths and needs (powell, 2019). Given that AI enables more flexibility of how students interact with learning resources compared to older edtech designs, the breadth of possible adaptations to address student needs and strengths is greater than in the past.
- **Supporting educators' skills, knowledge and implementation**. For example, this might include performing well in explaining student evidence to educators and providing a rationale for recommendations; aligning to an educator's curricular resources, such as their existing high quality instructional materials (HQIM); supporting educators' role as instructional leader; and creating better opportunities for educators to increase their professional knowledge as they go. Formative assessment, overall, is a long-standing process in education; thus, much is known

about how to support educators skills, knowledge and implementation. We expect proposals to be grounded in this body of knowledge.

Over the course of four years and multiple RFPs, we aim to build a collection of public goods that enables breakthroughs in the design, improvement, and implementation of AI-enabled, research-based formative assessment practices to support educators and improve student success. Within this vision, one opportunity for impact is related to **validity**: this program could fund deliverables that support product and service developers in establishing validity of formative assessments with respect to targeted universalism. This includes, for example, producing resources that support technical users in establishing fairness given the potential for bias. There are many validity concerns and we are looking for ways that especially address aspects of validity that lead to immediate practical applications for a wide range of product and service developers.

We encourage attention to the section associated with this RFP, **Formative Assessment and Public Goods for AI in Education**, for more insight into the expectations of this RFP. For more information, closely read the **Evaluation Criteria 1: Significance** in the **Proposal Guidance** section below.

Defining Public Goods

We are investing in public goods that can be adopted by a technical user (e.g., a learning engineer, data scientist, model developer, educational application developer, etc.). The public goods must be modular, foundational building blocks that lead to developing or improving educational applications for formative assessment. We anticipate public datasets being valuable in their own right, as well as public datasets that provide a foundation for creating benchmarks or improving models.

Three kinds of public goods are possible. Proposers should consider the modest amounts of funding in this first RFP as they decide what kind(s) of public good to focus on. The significance section can explain how the focal public good(s) could lead to other public goods, e.g., a dataset can be used by others to develop a benchmark or fine-tune a model, or could be used as a basis for creating additional synthetic data, etc. The three types are:

1. **AI Datasets**: Collections of data curated and structured specifically to train, fine-tune, support, or evaluate applications of AI toward our vision. Examples may include transcripts, instructional materials, student work, student drawings, etc. We emphasize our interest in datasets explicitly designed to enable benchmarking and evaluation workflows, not only model pre-training and fine-tuning. Datasets should be appropriately annotated for use and support the development of further annotation. We also anticipate supporting advanced workflows that extend layers

of annotation (e.g., additional human raters) and/or generate additional synthetic data. With regard to validity concerns, we welcome datasets that could be used to establish one or more aspects of validity. We will only fund projects that collect data responsibly, ensure that student privacy is respected, and enable responsible AI developers to evaluate if they are meeting target benchmarks and serving students.

2. **Evaluating AI Performance - Benchmarks and Models as Evaluators**: Standardized tools, tasks, metrics, and methods used to evaluate how well AI models (or AI-powered education technologies) perform in education-specific contexts, based on metrics like accuracy, fairness, relevance to learning goals, fitness for intended purpose, or ability to predict performance on standardized and widely-used learning measures. Benchmarks are not limited to validity; in one example, benchmarking one or more aspects of how well a formative assessment process supports an educator is important. Overall, a growing collection of public-good metrics will drive improvements over time to ensure that AI is safe, valid, meaningful, and attractive to educators. We are especially interested in component dimensions of evaluation where comprehensive measures might be misleading (in particular, we do not expect a monolithic validity benchmark). For example, proposers may suggest a modular collection of related benchmarks or models as evaluators (e.g., multiple dimensions for feedback quality or another aspect of formative assessment) that others can combine and test, rather than a single omnibus rating. Benchmark projects will be expected to result in standard tasks + metrics + example data aligned to relevant dimensions (not a single pass/fail).

3. **Advancing AI Performance - Training Models**: Machine learning algorithms that process K-12 educational data and perform tasks specific to evidence-based teaching and learning, such as identifying effective instructional practices or generating real-time feedback for students. By sharing models, innovators, researchers, developers, and others can build on the successful research of others. The program is interested in funding model-related public goods that are designed for integration into existing formative assessment applications, allow for adaptation and modification, improve domain representations (e.g., structured math representations), and produce model-building tools, pipelines, or processes that remain useful even as frontier models change. Public goods that can run locally or at low-cost or that a technical user can more easily modify for a specific educational context could be attractive (and we are aware that real educational products and services are using multiple models, so lowering costs for a targeted purpose could be valuable even where an overall product continues to use other higher-cost models as well). Models may also be a technique to provide the value of a large dataset as a public good without releasing the data itself.

We are investing in public goods that are useful to many organizations that create products and services for K-12 education. Recipients of funding will be required to license their funded public goods via a suitable license at least as permissive as Creative Commons Attribution (CC-BY-4.0) for content-like resources or Apache 2.0 for code-like resources. In all cases, recipients will not be able to exclude any type of organization from using the public good for an allowable purpose (e.g., to improve teaching and learning).

This approach does not require applicants to use the same approach to licensing for unrelated IP used parallel to these systems, nor for the release of raw data used to create these systems. In all cases, privacy, confidentiality, and ethical data standards will be applied.

For more information, see the section that discusses **Evaluation Criteria 4: Release and Dissemination Plan** in the **Proposal Guidance** section below.
Note: Proposals to develop proprietary product enhancements will be returned without review.

Assets and Capacity to Enable Rapid Progress with Privacy and other Protections

We will aim to provide funding quickly. Projects should be able to start close to June 1, 2026. Correspondingly, we are seeking teams with capacity and readiness to produce a public good. For example, we seek teams with existing datasets that require a modest amount of enhancement to become high-value public goods. We also seek teams that already have exercised the necessary processes to produce a public asset, which could include data de-identification tools and annotations, or more generally, data sharing agreements, MOUs, and privacy policies that are ready to go. Recognizing that high trust is essential for the future of AI in education, the program will only fund efforts that agree to and can confidently deliver on high standards for data safety and privacy. Assets and capabilities that enable a team to enact a targeted universalism approach will be important. Overall, we seek teams that can build on prior work to enhance or create a public good expediently yet carefully.

Grant Opportunity and Expectations

Eligibility

This opportunity is open to any U.S.-based applicants, including edtech companies, for-profit organizations, non-profit organizations or school districts. Only one application per lead organization will be accepted. Except that for universities and other organizations > 2000 staff, one application per department will be accepted. We encourage

applications that involve educators or field-facing organizations that directly work with educational practitioners. *Note: Partnerships among organizations are encouraged.*

Tracks

A responsive proposal must address one of the following tracks:

- ➔ **Track 1: Proof of Concept.** Low-cost prototypes or other initial efforts that can provide evidence that the approach could result in a more broadly available public good. This track is appropriate for making initial progress on an uncertain yet highly innovative approach, and should focus on tackling key challenges rather than producing a comprehensive, large-scale public good.
- ➔ **Track 2: Enhancing an Existing Asset.** Short-term efforts that leverage available datasets, models or benchmarks to rapidly produce a public good, e.g, within 6-12 months. This track is appropriate for projects that conduct additional technical work to enhance an already-public dataset, to integrate new data sources into a higher-value database, or to convert an existing dataset into a resource that can be more broadly used. Released assets must meet FAIR and other guidelines to enable uptake.

In either track, an aspect of a proposal could involve releasing resources that make it easier to protect privacy, enhance security or otherwise reduce the potential for harm within a set of formative assessment-related public goods. In general, it will be considered valuable if a project both produces a public good aligned with our vision, but also releases tools that could support further enhancing that public good or creating other, similar public goods.

Budget and Period of Performance

This solicitation invites proposals for $50,000 to $250,000 for a 6-12 month period of performance. We encourage applicants to request the funds needed to do the work, not the maximum allowed; doing so will increase the number of awards that are possible and cost-efficiency will be considered in proposal review. We anticipate making 4-8 awards.

The recommended starting date is June 1, 2026.

Allowable costs include:

- Staff time, which can include buy-out time for educators who work on the project
- Consulting fees or stipends to experts and advisors (who may be located internationally)
- Equipment and computational resources, (including devices, internet access, privacy-compliant analytical applications, computing power, and storage)
- Travel where necessary to producing or disseminating the public goods

- Costs of Institutional Review Board (IRB) review, privacy oversight, legal support for data sharing agreements, etc.
- Software or tools that are essential and allocable to the public good being developed
- Subawards or subcontractors performing project-specific tasks
  Translation or accessibility services to support equity and access
- Institutional overhead *up to a maximum of 15%*

Unallowable costs include:

- Meals, snacks, or alcoholic beverages
- Entertainment, social, or networking events
- Lobbying or advocacy related expenses
- Pre-award costs
- International travel not directly tied to project deliverables
- Expenses that are included in institution overhead such as:
  - General office supplies not directly allocable to the project
  - Administrative or clerical salaries not directly tied to the project
  - Capital expenditures (e.g., buildings, renovations)
  - Equipment purchases over $5,000 unless pre-approved and justified

This RFP is issued as part of a multi-year, $26 million program. Over the next three years, the program will issue additional RFPs as well as more grants. These may have additional tracks and different funding levels.

When and Where to Submit

Submissions will be accepted until March 8, 2026, midnight in Pacific Standard Time (PST), on this website. You will be asked to register in order to access the application.

What to Submit

Your submission should include a completed form including an abstract, the required PDF attachment, and a separate spreadsheet for the budget. The abstract (300 word limit) will be used to select appropriate peer reviewers. Specifically, the narrative and budget should contain:

The Narrative PDF:

- Please include your abstract in this PDF as well as the form field.
- A description of your project and narrative addressing the topics above (5000 word limit). Proposers should note that while this RFP provides detailed guidance on how proposals will be evaluated, many of the guidelines can be addressed concisely. A separate paragraph for each guideline is not required.

- Supplemental materials (up to 5 pages) can include prior work or early prototypes, such as initial data samples, mock-ups, or demo items
- Resumes or CVs for the project director and any other key staff
- If additional organizations or consultants are named in the project description, short letters that they have reviewed their role and are willing to serve (e.g. similar to NSF Letters of Commitment). These letters are required for documentation purposes only and will not be reviewed by peer reviewers.
- References and Citations, not included in the word count
- Please aim for legibility and follow standard proposal formatting such as: 12pt font size for text, at least 1.15 line spacing, a minimum of 10pt font for images/tables, and 1" margins.
- With the file name convention **PI LAST NAME_K12AI_RFP1_Narrative.PDF**

The Budget document will be submitted as a .XLS or .XLSX file and should include:

- Budget in the template provided, along with justification.
- Link to forced copy of Google Sheet
- Link to download .XLSX file
- Download a copy of the template and save your version with the file name convention **PI LAST NAME_K12AI_RFP1_BUDGET.XLS**

Evaluation Criteria

Proposals will be reviewed by both peer reviewers and the program team on four equally important criteria:

- **Significance**: WHY the project is important to K-12 U.S students and their teachers and to the market that develops services and products for this audience.
- **Assets and Capabilities:** WHAT and WHO will enable project success.
- **Project Workplan**: HOW the project will be completed
- **Release and Dissemination:** OUTPUTS of the project.

An FAQ will be available on the **K-12 AI Infrastructure Program website**. We will also offer office hours to respond to questions about the RFP.

Proposal Guidance

For ease of reviewing, proposals should have clear sections corresponding to the evaluation criteria and following the purpose described above for each section (i.e, Why, What and Who, How, Outputs). There is no recommended or required length by section, but as described above, the total length is limited to 5000 words.

The text below will be read by reviewers as guidance for their ratings, thus writing to this guidance will be valued. Notwithstanding this proposal guidance, submissions may include additional or different information to make the best case for their effort. We provide illustrative examples in the text boxes below. Examples are **NOT** requirements; we welcome proposals that differ from these examples.

Evaluation Criteria 1: Significance

Proposals must target the program's formative assessment vision, but can save space by assuming that reviewers understand the broad rationale for formative assessment.

Focus within Formative Assessment. The first page must identify the specific unmet needs, gaps, or opportunities within formative assessment that will be addressed, and provide the rationale for why the team will be able to make rapid progress in strongly addressing that specific need, gap, or opportunity. Why will these specific improvements to formative assessment move the needle? What will be the scope of the proposed public good (size, modality, subject area, grade levels)?

> For example, a proposal might explain how a new model would enable dialogue with students that better elicits their knowledge; or how a benchmark for explainability of formative assessment recommendations would increase the value to educators of engaging in formative assessment. In another example, could the public good help developers establish one or more aspects of validity for their product or service?

***Note:** The remaining Significance guidelines can be addressed in any order.*

Operationalizing Research. What learning sciences or other research-based concepts, methods, or insights will be incorporated into the public good?

> For example, a public good might establish benchmarks for the quality with which formative assessment interactions engage student prior knowledge, use visual images, simulations, or other artifacts to draw out student mental models, or identify misconceptions as well as strengths. Or a public good might establish benchmarks for the quality with which an instructional planning process is meaningfully differentiated for the learner variability uncovered by evidence of recent student work and learner variability captured in background characteristics of the learners. A benchmark might enable a developer to quantify specific dimensions of validity, too, per the discussion above.

Targeted Universalism. How will the advance enabled by the proposed public good(s) enable success for every student? Why will it make an important difference for targeted populations of students who otherwise might be underserved?

> For example, a speech dataset where teachers interview students about their prior knowledge

would be particularly valuable if the speakers have linguistic backgrounds or ways of speaking that vary from what is available in existing datasets or that represent realistic variability among learners. Likewise, a benchmark for the quality of implementation of a learning sciences principle should enable implementations that help every student, but also attend to learner variability so that students with specific disabilities, backgrounds, existing strengths, or preferences are well-served.

Type(s) of Public Good(s). If the effort will produce a benchmark, why will that drive the field forward and lead to impacts for millions of students? Likewise, why is producing a dataset, model, or an alternate public good the right choice? What existing resources exist and what potential benefits will this work provide?

For speech datasets, one might argue that this is how automated speech recognition progress occurs. For benchmarks, one might show that they will call attention to a specific kind of previously unnoticed but highly important variability across existing AI models (e.g. a kind of bias that interferes with validity), and this might increase attention to improvements that improve formative assessment for a population of at-risk students.

Potential for Wide Adoption and Large-Scale Student Impacts. Which technical end-users are likely to adopt the public good, and what benefits and cost advantages will they realize? Is there evidence of demand or likely usage? Why can we expect that the set of possible adopters will actually use the proposed public good(s) in offerings that rapidly scale to large numbers of students?

For example, a knowledge graph that can better align the resources in a very popular high quality instructional material (HQIM) to formative assessment decision processes might be adopted by many supplementary providers who support educators that use the HQIM as their core curriculum. Or a hyperscaler might incorporate a dataset of examples of how expert teachers elicit student strategies or reasoning to adapt their next instructional move to improve either their overall foundation model, or a more specific education-oriented model such as a "learning" or "study" mode. Obviously, if a hyperscaler chooses to incorporate a public good in their outputs, this improvement will become widely available.

Evaluation Criteria 2: Assets and Capabilities

For Track 1, describe the existing assets that can empower rapid progress on the proof-of-concept effort.

For Track 2, describe the **existing asset(s)** that will provide the basis for a public good. Prompts that might be helpful in creating your description include:

- What does the asset provide? Describe the existing asset.
- What is the current usage of the asset?

- Are there existing evaluations that establish the impact of the asset towards improving K-12 education?
- What makes it appropriate for the proposed public good?

  For example, an existing service may have collected a lot of data that reveals an important aspect of learner variability that could be addressed by applying learning science insights and may have seen that they are able to use the data to make a portion of the necessary improvements. Because they lack the expertise or funding to address all the kinds of learner variation present, they will release it as a public good, so that others may solve these challenges.
- How does the asset represent the targeted universalism goals described in your Significance?
- What tools, processes, and workflows exist to support the work?
- What equipment, compute, travel, facilities and other resources would you need to conduct your project successfully? Will you have access to these resources through existing means or through use of grant funds?
- What legal agreements and policies establish the rights to release the public good?

  For example, an existing data sharing agreement can be briefly summarized and later provided to the program team for pre-award review. If the existing licensing rights are compatible with CC BY or Apache 2.0 licensing that can be stated. Or will re-consent or other processes be necessary to obtain rights?
- What types and sensitivity of data will you collect, and how will it be managed responsibly?

  Who will be responsible for data management, and what infrastructure do you have in place or will you need to build? If you are proposing to use existing data, state for what purpose the data were originally collected.

For both tracks, describe the sensitivity of data that will be used during the period of performance, and the data infrastructure that will be used along with any standards, certifications, etc.

For both tracks, discuss the **expertise** that will enable successful completion of the work as well as any available infrastructure facilities that are available to support the work.

- To what extent does your team have all the expertise necessary to execute the project successfully?

  For example, in machine learning, learning sciences, data science, and educational practice.
- Who has the expertise and experiences to enact the Targeted Universalism focus you described in the Significance section?
- Who has the expertise and experiences to manage any sensitive data the project will use?

- Who has expertise or experience in evaluating the education-specific value of the intended work?
- What depth of experience does your team have in producing public goods or education-focused AI resources similar to datasets, models, and benchmarks?

We encourage applicants to consider how the perspectives of educators, learners and other impacted communities are represented on your team. We encourage integrating a learning scientist with expertise on the learning science concepts that will be used, especially to ensure that those concepts are handled with fidelity.

Applicants may submit Supplemental Materials with any prior work or early prototypes, such as initial data samples, mock-ups, or demo items. They may also describe the essence of existing data sharing, IRB or other legal agreements that could be later shared with the program team for detailed review.

Evaluation Criteria 3: Project Workplan

The overarching questions for this section is: What is the justification and evidence that the workplan will confidently and safely produce the public good(s) described in the Significance section, building on what is available as described in the Assets and Capabilities section?

The workplan may begin by clearly addressing objectives: what needs to be done to produce the significance of the public goods using the assets and capabilities, both of which have already been described.

> For example, a Track 1 proposal would describe a small set of objectives for what the proof of concept will demonstrate or establish. A Track 2 proposal might describe the additional documentation, meta-data, dataset merges, etc. that are required to enhance the existing assets to be ready for release as a public good.

To make the argument for the workplan, proposers may cover aspects such as these in any order or combination:

- **Kickoff:** How a rapid start will be achieved once funding is available.
- **Phases, Milestones, Timelines:** Steps, phases, or workflows, showing how the work will proceed. Is adequate time available for each step? A chart is acceptable.
- **Centering:** How the targeted population(s) will be centered in the work.
- **Contingencies:** The most important contingencies for additional known risks and how they can be handled.
- **Roles:** For major responsibilities, is it clear who will do what? In particular, describe data management roles.
- **Data Security, Privacy Protection, Quality Assurance:** Funded projects must anticipate working with the K-12 AI Program Team during the period of

performance to complete a data management plan and participating in data governance review and audits

Note: The workplan can end with preparing "release candidates" of public goods; final quality assurance process can be described in the next section.

Evaluation Criteria 4: Release and Dissemination Plan

Track 1 proposals should release a public good (see info for Track 2, below), but the public good will be a prototype or proof of concept and therefore will be limited in scope. An expected deliverable for a Track 1 proposal is also a memo and briefing to the program team that explains what further work would lead to fully realized public good as further funding may be available after the conclusion of the Track 1 work.

Track 2 proposals require a more thorough release and dissemination plan around a completed public good. The program team will collaborate with awarded projects on release and dissemination. At a minimum, public goods will be hosted in the program's digital repository, which will be available at no-cost without restriction. Beyond releasing public goods, we seek to publicize and support the public goods so that they become widely known, adopted, and used to support students and educators at scale.

Public goods developed in either track must respond to FAIR Principles (Findable, Accessible, Interoperable, Re-usable). The minimum required alignment will be accomplished by working with the K-12 AI Infrastructure Program team to deposit the public goods in the Program's repository, with a DOI, download capability and appropriate documentation. Teams can propose to go beyond this minimum, especially where this would encourage greater adoption of the public goods.

The Release plan necessarily includes:

- **Quality Assurance.** Describe how the project team anticipates working with the program team to review and approve public goods prior to release, to ensure quality and protections for privacy, data security, and other safety concerns.
- **Licensing Terms.** The program team has selected Creative Commons (by Attribution) as the recommended license for datasets and knowledge products (e.g., technical reports, research results, technical reports) and Apache v2.0 for software and code (e.g. evaluations, models, applications). Other licenses may be negotiated during the pre-award phase, but a persistent requirement is that the resources are available for commercial or non-commercial use.
- **Limitations and restrictions** based on data usage agreements, IRBs or ethical considerations.

  For example, the provenance of data may require that it be licensed only for a specific field of use, e.g. K-12 instructional uses and not for consumer-oriented uses. If there are

limitations on use, e.g. cannot be used to train a model, then clarify those limitations.

- **Evaluations.** The project team will work with the Program's evaluation team to produce compelling evidence of the value of the public good, compared to existing practice. In support of this, the proposal can describe how they anticipate establishing fitness for intended use, known limitations and demographic bias considerations (e.g., reporting bias and fairness metrics), validity measurements, and indications that the public good(s) can advance the field on desirable metrics.
- **Documentation and Supports.** The team can describe the resources that will accompany the public good and that will make adoption and use more likely.

The identification of existing baseline performance (by humans or automated tools), benchmarks or existing evaluations to be used, and other criteria are especially helpful. We also encourage the use of models, sample code for using the public good, demonstrations of the improvement produced by employing the public good within the awardee's own context, etc.

Aspects of a compelling dissemination plan may include:

- **Possible Data Competitions.** To engage technical users across many organizations, the program intends to host data competitions around public goods. It would be valuable for teams to produce one or more designs for a data competition as the proposed project moves into its dissemination phase.
- **Broad Promotion.** The team can describe how they can directly promote the public good to an audience, such as through a webinar, conference presentation, targeted emails or posts in relevant forums, etc.
- **Specific Relationships.** The team may have relationships to a hyperscaler, education technology standards organization, or other partner whose incorporation of the public good may lead to large scale impacts on student success. If so, describe how those relationships could be activated.

References

powell, j. a., Menendian, S., & Ake, W. (2019). Targeted universalism: Policy & practice [Primer]. Othering & Belonging Institute, University of California, Berkeley. https://belonging.berkeley.edu/sites/default/files/targeted_universalism_primer.pdf

Formative Assessment and Public Goods for AI in Education

Author: K-12 AI Infrastructure Program Team

Definition

Formative assessment is a key aspect of many applications of AI to education. Through production of public goods (as defined in the *Request for Proposals*), the K-12 AI Infrastructure program seeks to aid developers of AI-enabled products and services. To ground the meaning of formative assessment, we start by defining it without respect to technology:

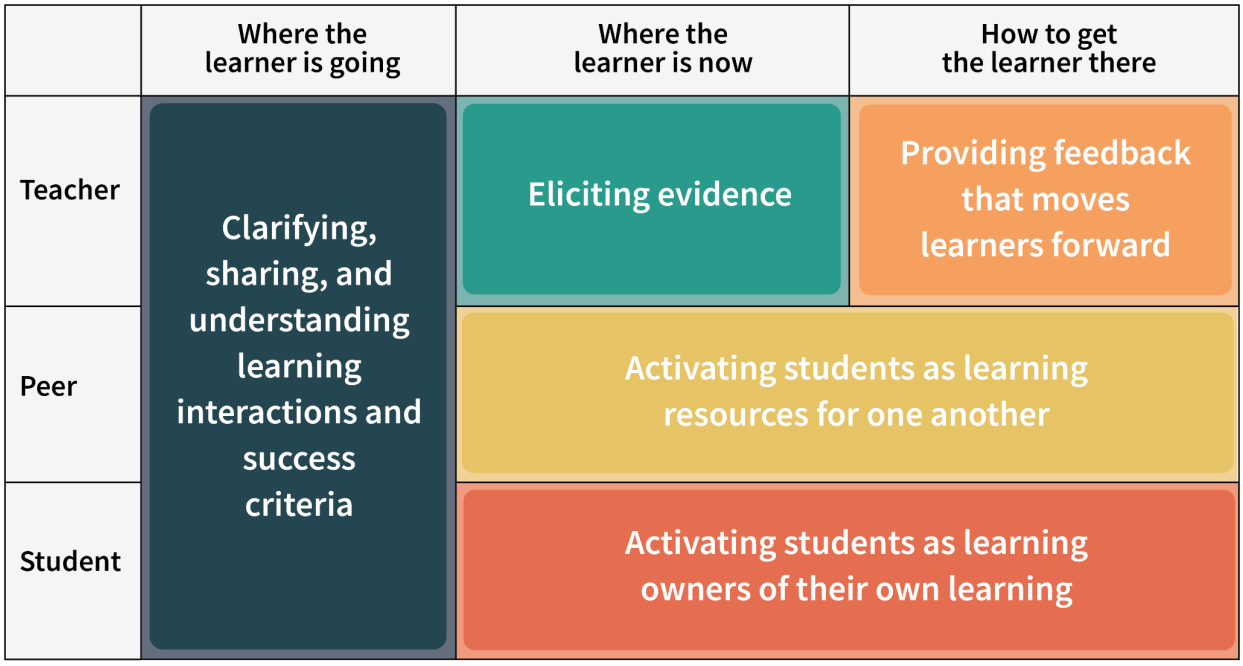
> "Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited." (Black & Wiliam, 2009, p. 9)

In practice, these decisions refer to the concrete instructional choices that teachers, learners, or peers make in response to evidence of learning. These decisions can range from immediate, in-the-moment adjustments such as rephrasing a question, offering a hint, or selecting a different example, to longer-term planning choices, such as revisiting a concept in the next lesson or modifying the pace of instruction. The core idea is that evidence is most useful when it changes what happens next for learners, and the changes increase learning.

Wiliam later developed the image in Figure 1, which usefully clarifies the realm of formative assessment as including of teacher, peer, and learner roles and requiring goal clarification, eliciting evidence, and planning next steps. Below, we will elaborate on the many opportunities for AI to enhance this image

Importantly, formative assessment is not limited to testing students in the conventional sense, e.g. a quiz. It can be based on classroom observations, discussions, and examining student work. Teachers are naturally involved in multiple ways: directly interacting with students, adapting instructional resources for student use, and stimulating students' active learning.

Figure 1: **Unpacking Formative Assessment.**



Source. Wiliam, D. (2025). *Formative assessment in an AI world* (conference presentation). Flourishing Learners Conference, Melbourne, Australia, Sept 8–9, 2025). Used with author's permission.

Similarly, "evidence" in formative assessment extends beyond academic performance to include student effort, motivation, social engagement, and other relevant strengths and assets. Paying attention to these dimensions helps teachers understand not just what students know, but how they are approaching their learning. This information can shape both instructional decisions and how teachers communicate with students.

However, gathering evidence alone is not sufficient, and it's only formative assessment if it leads to improved instructional decision-making. This can involve decisions about how to interpret the evidence, what to talk about with students, how to talk about it, and what instructional resources and activities to use next. These resources and activities can be generated or adapted to respond to the evidence.

Formative assessment is applicable across K-12 grade levels and subject matters. In the context of AI, this means that technologies such as computer vision, automatic speech recognition, and dialogue analysis can support evidence gathering, while AI-powered tools can assist with planning, recommending resources, and adapting instruction—expanding the possibilities for formative assessment at scale.

Efficacy

Many research syntheses and meta-analyses support the claim that Formative Assessment is efficacious (e.g., Klute et al., 2017). Many educators are aware that John Hattie (2009; 2012) lists feedback and formative assessment among the most effective instructional processes. And yet, like any educational intervention, the benefits of formative assessment depend on how well it is enacted (Kingston & Nash, 2011). Quality varies, and so do impacts.

The seminal review by Black and Wiliam (1998) examined more than 250 studies and found that formative assessment can lead to substantial learning gains. However, later research using more rigorous methods found smaller, but still meaningful, benefits. Kingston and Nash (2011) reviewed K-12 studies and found that the positive effects varied by subject, with English language arts showing the strongest benefits, followed by mathematics and science. More recent reviews confirm these patterns: Lee et al. (2020) and Jiang et al. (2024) both found consistent positive effects on student learning across dozens of studies in U.S. and international settings.

How formative assessment is delivered also matters. Graham et al. (2015) found that in writing instruction, feedback from teachers produced the strongest improvements, followed by student self-assessment, peer feedback, and computer-based feedback. A comprehensive review by Lipnevich et al. (2024) synthesizing 13 prior reviews concluded that formative assessment consistently helps students learn, with no studies showing negative effects. Overall, the research indicates that well-implemented formative assessment reliably improves student outcomes, though results depend on how well it is carried out, the subject area, and the specific practices used.

Validity and Reliability

Validity means that the formative assessment, when used for its intended purpose, provides a high quality of evidence of student learning and enables adapting instruction in ways that benefit student learning (AERA, APA, & NCME, 2014). Validity is not a property of an instrument alone but of the *interpretation-and-use argument* that links observed student work to instructional decisions (Cronbach & Meehl, 1955; Kane, 2006). Accordingly, developers should provide validity showing that the formative assessment works for its intended purpose. At a minimum, this evidence should show that: (1) the formative assessments are well aligned to the target knowledge and skills; (2) the formative assessments actually elicit the kinds of thinking and strategies the assessment is intended to measure; (3) scoring and reporting are consistent and accurate enough to support the kinds of instructional decisions teachers will make; (4) results relate to other credible measures in expected ways; and (5) the formative assessments supports beneficial and fair instructional actions for different groups of students (Messick, 1989;

AERA, APA, & NCME, 2014). Because instructional decisions depend on the stability of evidence, developers should also report the reliability/precision of scores or classifications, at a level appropriate to the intended formative decisions (AERA, APA, & NCME, 2014).

Teachers

There is substantial literature discussing the teacher professional development required for strong classroom implementations of formative assessment (e.g., Wiliam, 2018). One obvious issue is reducing the time or burden on teachers to gather evidence and adapt their instructional plans. And yet, when teachers offload formative assessment to 1:1 technologies that are directly used by students, the lack of coordination and alignment with teacher-led instruction can emerge as a major challenge. It can also be challenging to make sense of student thinking (to interpret the evidence) and to adapt instruction in ways that remain coherent with underlying curricula and yet respond to students' specific strengths and needs. Also note that formative assessment is often integrated into more comprehensive programs of teacher professional development. For example, Cognitively-Guided Instruction (CGI) is a well-known program in mathematics; it focuses on making sense of and building on students' mathematical strategies and problem solving approaches. In CGI, formative assessment is a component of a broader program of teacher professional development for mathematics educators.

Technology

Technology can and often is used to support formative assessment. Rigorous studies have established that this use of technology can be efficacious (Roschelle et al., 2016), with some syntheses claiming greater effect sizes for formative assessment when it is supported by technology (van der Kleij et al., 2015).

In one example, Roschelle et al. (2016), a formative assessment intervention was defined using ASSISTments software. In middle school math classrooms, teachers assigned homework problems to students in ASSISTments, and students received feedback, hints and guidance as they worked on the assignments at home. Teachers' workload decreased because ASSISTments provided the results of the student homework in a neatly organized table. Teachers' practice changed because they could focus on specific homework problems and wrong answers in classroom discussions (Fairman, Feng, & Roschelle, 2025). In other words, there was a modification to teachers' next steps in instruction, teachers spent more time on fewer problems, and targeted their discussion to the evidence of what support students needed. A significant and meaningful effect size was observed, and the effect was stronger for students who were struggling in mathematics (Murphy et al., 2020). In another study, a longer term effect was observed, as well (Feng, Huang, & Collins, 2023).

The ASSISTments example is useful because it is a case where technology both supported students directly and helped their teacher. For the teacher, it helped in their preparation of formative assessment assignments, in grading, and in interpretation leading to instructional decision making. The study also demonstrated one aspect of targeted universalism: students with weaker prior performance learned more, but students with stronger performance also gained.

Learning Sciences Concepts

A wide range of established learning science concepts can be used to guide formative assessment. These research-based principles describe what effective teachers do when gathering and responding to evidence of student learning. They can be organized around three core activities (Table 1).

Table 1. Foundational Learning Sciences Concepts

Eliciting evidence	• Eliciting and making sense of students prior knowledge (Ausubel, 1968; National Research Council, 2000) • Understanding students' reasoning and strategies (Schoenfield, 1985) • Distinguishing productive struggle from wheel-spinning, guessing, or other less productive behaviors (Beck & Gong, 2013)
Interpreting evidence	• Addressing student misconceptions (Chi, 2005; Smith et al., 1994) • Engaging students' funds of knowledge (or assets) in instructional planning (Moll et al., 1992)
Acting on evidence	• Giving high quality feedback (Shute, 2008) • Using multiple representations (Ainsworth, 2006) • Knowledge integration (Linn, 2006) • Supporting students self-regulated learning (Zimmerman, 2002)

Evidence-Centered Design (ECD; Mislevy et al., 2003) is a research-based framework that can be used to systematically design, develop and validate assessments. In K-12 formative assessment, the goal is to elicit evidence to make immediate instructional adjustments and guide student next steps (Black & Wiliam, 1998; Heritage, 2010). Evidence-based design applies ECD's claim–evidence–task structure as a practical "blueprint": specify the learning claim (what proficiency looks like), define what counts as

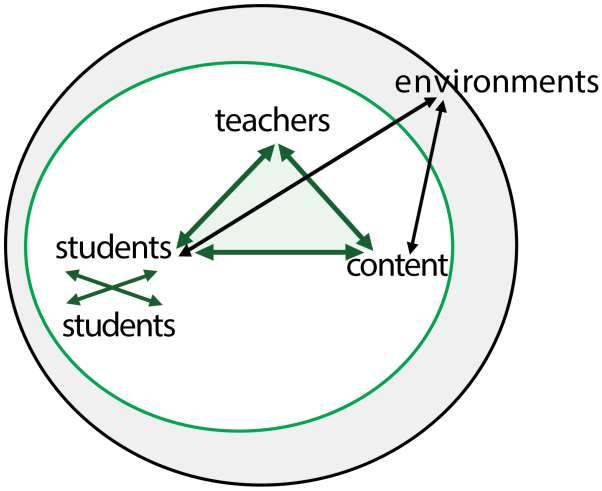
evidence (e.g., work products, explanations, error patterns), and design scaffolds and scoring guides that reliably elicit and interpret that evidence during instruction (Mislevy et al., 2003). Crucially, ECD also makes the inference explicit by articulating why the selected evidence warrants the claim and by specifying decision rules for how different patterns of evidence trigger a particular instructional response (e.g., reteach, extend, scaffold). This strengthens formative assessment by making classroom checks (e.g., questioning routines, exit tickets, performance tasks) intentionally diagnostic and by supporting feedback that is timely, specific, and usable so the information collected is actionable for both instructional adjustments and student revisions (Shute, 2008; Wiliam, 2011). When shared as clear success criteria, the same evidence and decision rules also support students' self-monitoring and revision during learning, and not just teacher action.

AI and the Future

In practice, formative assessment is central to most applications of AI in education. In AI tutoring, for example, the system examines evidence of what a student knows or is struggling with, plans how to respond, and delivers instruction while monitoring for improvement. While "personalized instruction" is often a vague aspiration, formative assessment offers a more precise framework for how instruction adapts to student needs based on evidence. Of course, responsible use of AI for formative assessment requires establishing that it can perform these functions with validity.

It can be useful to think about AI's role in the context of the Instructional Triangle (Cohen, Raudenbush, & Ball, 2003), which analyzes instruction as occurring among teachers, students, and content in an environment or context. AI can potentially participate in the triangle in multiple roles, for example as a tutor, as an assistant to a teacher, as a teachable agent (cite) that allows students to express what they know, as a support for students to collaborate with each other, or in adapting content.

Figure 2. The Instructional Triangle (Cohen, et al., 2003)



The Instructional Triangle (Cohen, Raudenbush, & Ball, 2003) analyzes instruction as interactions among teachers, students, and content within an environment. AI can participate at each point of this triangle, creating opportunities for formative assessment support:

Teacher–Student interactions

- Clarifying instructional goals (e.g., as broadly given in curricular resources) and expressing what success looks like
- Reducing teacher workload associated with evidence gathering (e.g., interviewing students, analyzing their drawings for misconceptions or strategies, helping narrow down the focus for working with students)
- Strengthening teachers' insights about their students (e.g., making sense of student reasoning, misconceptions, strengths, learning variability)
- Talking with students and looking at their work to better understand students
- Planning activities, discussions, worksheets, etc. to tune support for student learning based on the evidence
- Using formative assessment as an opportunity for teacher coaching

Student-Content interactions

- Providing high quality feedback and guidance to students
- Planning appropriate next instructional steps for students based on the evidence, e.g., optimizing their practice on a skill they are learning or retaining

Student-Student interactions

- Using evidence to create small groups around a specific learning need, and regrouping as needed as understanding changes.
- Designing collaborative small group activities that require the participation of individual students

Teacher-Content interactions

- Understanding the design of their high-quality instructional materials as they reason about their students
- Choosing or adapting their high quality instructional materials based on the evidence

Environment/Context

- Connecting formative assessments with other factors in the classroom or other educational setting, which can relate to school-wide strategies or policies (e.g., regarding homework or instructional design), other benchmark, periodic or end-of-the-year assessments, teachers' preferred pedagogies, available or recommended instructional resources, understanding of students' family or community setting

The K-12 Infrastructure Program, in its first RFP, asks for proposals that produce public goods that advance formative assessment while incorporating one or more of the following:

1. Multimodality, including using spoken language and making sense of student-drawn images.
2. Learning sciences, operationalizing known learning science principles that are supportive of formative assessment
3. Educator supportiveness, including features like explainability, support for teacher co-design, support for teacher learning, etc.

See the *Request for Proposals* for more information.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart & Winston.

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (pp. 431–440). Springer. https://doi.org/10.1007/978-3-642-39112-5_44

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). *Developing the theory of formative assessment*. *Educational Assessment, Evaluation and Accountability, 21*, 5–31. https://doi.org/10.1007/s11092-008-9068-5

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. Journal of the Learning Sciences, 14(2), 161–199. https://doi.org/10.1207/s15327809jls1402_1

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. Educational Evaluation and Policy Analysis, 25(2), 119–142. https://doi.org/10.3102/01623737025002119

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302. https://doi.org/10.1037/h0040957

Fairman, J. C., Feng, M., & Roschelle, J. (2025). *Teachers' use of learning analytics data from students' online math practice assignments to better focus instruction. Digital Experiences in Mathematics Education*. https://doi.org/10.1007/s40751-025-00170-3

Feng, M., Huang, C., & Collins, K. (2023). *Promising long term effects of ASSISTments online math homework support*. In N. Wang et al. (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2023): Late*

Breaking Results (pp. 212–217). Springer Nature. https://doi.org/10.1007/978-3-031-36336-8_32

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. The Elementary School Journal, 115(4), 523–547. https://doi.org/10.1086/681947

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom.* Corwin. https://doi.org/10.4135/9781452219493

Jiang, Y., Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2024). The impact of formative assessment on K-12 learning: A meta-analysis. Educational Research and Evaluation, 29(7–8), 423–450. https://doi.org/10.1080/13803611.2024.2363831

Kane, M. T. (2006). Validation. In R. B. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A review of the evidence (REL 2017–259).* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. Retrieved from http://ies.ed.gov/ncee/edlabs.

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K-12 education: A systematic review. Applied Measurement in Education, 33(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Linn, M. C. (2006). The Knowledge Integration Perspective on Learning and Instruction. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 243–264). Cambridge University Press.

Lipnevich, A. A., Guo, F., & Tay, L. (2024). A systematic review of meta-analyses on the impact of formative assessment on K-12 students' learning: Toward sustainable quality education. Sustainability, 16(17), 7826. https://doi.org/10.3390/su16177826

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 1*, i-29. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory Into Practice, 31*(2), 132–141. https://doi.org/10.1080/00405849209543534

Murphy, R., Roschelle, J., Feng, M., & Mason, C. A. (2020). *Investigating efficacy, moderators and mediators for an online mathematics homework intervention*. *Journal of Research on Educational Effectiveness, 13*(2), 235–270. https://doi.org/10.1080/19345747.2019.1710885

National Research Council. (2000). How people learn: Brain, mind, experience, and school: Expanded edition. National Academies Press. https://doi.org/10.17226/9853

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open, 2*(4), 1–12. https://doi.org/10.1177/2332858416673968

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Smith, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3*(2), 115–163. https://doi.org/10.1207/s15327809jls0302_1

van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Wiliam, D. (2018). Assessment for learning: meeting the challenge of implementation. *Assessment in Education: Principles, Policy & Practice, 25*(6), 682–685. https://doi.org/10.1080/0969594X.2017.1401526

Wiliam, D. (2025, September 8–9). *Formative assessment in an AI world* [Conference presentation]. *Flourishing Learners Conference*, Marvel Stadium, Melbourne, Australia.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2