Empowering Educators via Language Technology

Dorottya (Dora) Demszky, Jeffrey B. Bush, Sidney K. D'Mello, Jennifer Jacobs, Isabelle Hau, Heather Hill, Jing Liu, Susanna Loeb, Bethanie Maples, Kylie Peppler, Rhea Pokorny, Matthew Rascoff, Jenny Robinson, David Yeager, Laura Wentworth

Stanford University

Empowering Educators via Language Technology

Authors (except first author all are listed alphabetically by last name): Dorottya (Dora) Demszky[1], Jeffrey B. Bush[2], Sidney K. D'Mello[2], Jennifer Jacobs[2], Isabelle Hau[1], Heather Hill[3], Jing Liu[4], Susanna Loeb[1], Bethanie Maples[1], Kylie Peppler[5], Rhea Pokorny, Matthew Rascoff[1], Jenny Robinson[2], David Yeager[6], Laura Wentworth[7]

Affiliations:
[1] Stanford University
[2] University of Colorado, Boulder
[3] Harvard University
[4] University of Maryland, College Park
[5] University of California, Irvine
[6] University of Texas, Austin
[7] California Education Partners

Corresponding Author: Dora Demszky (ddemszky@stanford.edu), Assistant Professor at the Stanford Graduate School of Education

Acknowledgments: *This publication is supported by the Bill and Melinda Gates Foundation and Stanford Digital Education. D'Mello would like to acknowledge NSF DRL 2019805. Bush and Jacobs would like to acknowledge NSF DRL 2222647. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF. We thank Rhea Pokorny and Jenny Robinson for writing support. We are grateful to Hiep Ho for formatting the paper.*

Table of Contents

Introduction

By May 23, 2023, when academics and industry professionals met on the Stanford campus to discuss the future of natural language processing (NLP) in education, ChatGPT was already changing how students and teachers perceive learning goals and processes. Yet ChatGPT is only one example of NLP technology, which allows computers to process and produce human-like language, and that promises to bring extraordinary power to efforts to address persistent challenges in education. Dora Demszky, assistant professor in education data science at Stanford Graduate School of Education, who had assembled the group of academics, industry professionals, and people working in the education sector, highlighted the ambition that motivated the conference. “The goal is to help build and shape the field,” she told attendees.

As they introduced themselves, attendees expressed hope that increasingly sophisticated language technology could improve many facets of education, from targeted ideas like “enabling teachers to engage in learning protocols with a coach” (Adam Geller, founder and CEO of Edthena) to large-scale aims such as realizing the “potential for adaptive learning” (Sean Hobson, Chief Design Officer, Arizona State University). This excitement paired with apprehension; each vision had its own inherent risk. “I’m excited about the possible equity that can be achieved with one-to-one instruction but I’m worried that’s not going to be the case,” Stanford computer science graduate student Tolúlọpẹ́ Ògúnrẹ̀mí said.

Through visioning exercises, in-depth challenge discussions, and informal conversations, participants explored the potential of language technologies. This whitepaper attempts to represent, in broad form, the abundance of ideas that were articulated, while suggesting areas of agreement and tension. In this, we are inspired by and in dialogue with the insights issued by the U.S. Department of Education in its report on AI and the future of learning.[1] We hope that the specific points and large brushstrokes of our conversations at the conference can form a basis from which new discussions and partnerships can develop.

[1] Office of Educational Technology, U.S. Department of Education, “AI and the Future of Learning: Insights and Recommendations,” May 2023, accessed September 26, 2023, https://www2.ed.gov/documents/ai-report/ai-report.pdf; also available at https://tech.ed.gov.

Guiding Principles

Facilitators Dora Demszky and Heather Hill (professor at the Harvard Graduate School of Education), laid out four guiding principles, grounded in survey responses provided by conference participants in a pre-conference reflection exercise: beginning with equity, centering teachers' needs, promoting high-quality instruction, and building and informing educational theory. These principles echo many of the commitments discussed in a recent report by UNESCO.[2] Underlying all of these principles is the objective to have "AI in the loop, educators in charge" at every stage of technology development, constantly assessing and analyzing impacts in the classroom.

Begin with Equity

We define "beginning with equity" as prioritizing students, families, teachers, and schools who have historically been underserved by existing education systems. In the NLP context, this means designing technology for those groups and committing to making that technology accessible. Well-designed NLP should improve equity across classrooms by assisting under-resourced classrooms and helping teachers provide linguistically and culturally sensitive lessons, among other goals. However, tools implemented without equity at the forefront have the potential to exacerbate existing inequalities. Bias in existing datasets and modeling approaches, discriminatory implementation, and resource inequalities may hurt already underserved communities. Instead of addressing equity considerations in the later phases of a project, researchers and entrepreneurs should begin with them to drive processes, inputs, and assessments.

Center Teacher Needs

"Our [NLP] community has a shared vision of uplifting and empowering teachers," Jennifer Jacobs, associate research professor at CU Boulder, said while introducing herself. "I think we need to come together and say it." Interventions that require teachers to make additional large sacrifices in their classroom or personal lives will not be effective. In 2023, 58 percent of K-12 teachers reported workload-related

[2] UNESCO, "An Ed-Tech Tragedy?," 2023, summarized by Singer in The New York Times, https://www.nytimes.com/2023/09/06/technology/unesco-report-remote-learning-inequity.html.

stress and burnout, compared to 33 percent of working adults overall.[3] Adding in more technologies can exacerbate this burnout; one recent study found that a single teacher already accesses an average of 148 unique educational tools each year.[4] Given the overload of tools and information that teachers are expected to deploy, they may understandably resist new interventions, even potentially beneficial ones.[5] Developers must integrate teachers in their development process, ensuring that innovations are not only for teachers but also co-developed with teachers. This way, developers can ensure that tools map to teachers' real needs and work effectively with existing school infrastructure.

Promote High-Quality Instruction

Better teaching practices lead to better student outcomes.[6] Language technology has the potential to promote high-quality instruction by targeting a wide range of learning mechanisms: strong teacher-student relationships, rigorous and engaging curricular materials, and productive teacher-student interactions, among others. Rigorous evaluation is needed to identify the most effective approach for different contexts. There is broad agreement on what high-quality instruction looks like but researchers still find very uneven levels of quality across teachers, classrooms, and schools — an issue that has disproportionately negative effects on students of color. Minimizing this variation is an important goal.

Build and Inform Educational Theory

We want technologies to meaningfully expand what we know about how people learn, including what practices best support equitable opportunities for engagement, knowledge sharing, higher-level thinking, and skill development.

Researchers studying open questions in education would benefit from high-quality data and adaptive technology that can interpret that data at scale to yield robust

[3] Madeline Will, "What's Happening to Teacher Stress Levels," Education Week, June 21, 2023, accessed September 25, 2023, https://www.edweek.org/teaching-learning/whats-happening-to-teacher-stress-levels/2023/06

[4] Instructure, "EdTech top 40 report shares the latest on the usage of digital tools during the 2021-22 school year," August 24, 2022, https://www.instructure.com/resources/blog/edtech-top-40-report-shares-latest-usage-digital-tools-during-2021-22-school-year

[5] Alyson Klein, "Tech Fatigue Is Real for Teachers and Students. Here's How to Ease the Burden," Education Week, March 8, 2022, https://www.edweek.org/technology/tech-fatigue-is-real-for-teachers-and-students-heres-how-to-ease-the-burden/2022/03.

[6] Linda Darling-Hammond, "Teacher Quality and Student Achievement," Education Policy Analysis Archives 8 (2000): 1, https://doi.org/10.14507/epaa.v8n1.2000.

evidence. For example, understanding how teachers' instructional discourse affects students' learning outcomes has so far been bottlenecked by the lack of large, diverse datasets collected from controlled environments, thereby limiting causal relationships to be teased apart. Language technology can help address this gap by allowing us to i) collect large-scale transcript data via automated transcription, ii) provide scalable measurements of instructional practice, iii) provide personalized feedback to teachers to support their understanding and use of best practices and finally, iv) measure how specific instructional practices are related to outcomes of interest.

Just as it should be a priority for NLP technology to inform theory, whenever possible, it should also be informed by leading theories of learning. The design of technological tools should reflect existing, foundational research in the field of education. For example, rather than facilitating rote memorization, technologies should encourage students to engage deeply with content through creative problem-solving, collaboration, and in-depth exploration. Socially mediated, culturally responsive and collaborative theories of learning can be reflected by NLP technologies that uplift student contributions, promote divergent thinking and catalyze collaborative classroom interactions.

Strategic Directions

At the conference, participants broke off into pairs to discuss a series of questions about visions for the future. These exercises, meant to be generative and imaginative, yielded four main categories of educators' work that new innovations could support: routine teacher tasks, professional learning, curriculum development, and formative assessment.

Supporting Routine Teacher Tasks

"Teachers are being asked to do too much with too little," said Norma Ming, manager of research and evaluation at San Francisco Unified School District. A teacher's job does not end at instruction. NLP can "automate routine tasks for teachers such that teachers can focus on activities of higher value, such as building meaningful relationships with students." Automating tasks, especially ones that are time-intensive or repetitive, could alleviate teacher burnout.

NLP can streamline some daily tasks both inside and outside of the classroom. For example, tools could take attendance, freeing up the five minutes a teacher might use, or assist in email generation, sending automated homework reminders to students or a monthly "what we did in class" report to parents. AI could autograde multiple choice assessments and deliver grade summaries to teachers, so that teachers could spend less time checking answers and focus instead on instruction. There is potential, as well, for AI tools to grade non-multiple-choice assessments, though this would require rigorous teacher review of AI-generated scores and feedback.

Teachers can also use NLP to engage more deeply with student work — for instance, by consulting AI tools that might analyze that work and suggest how to assist students, another process that would require careful review on the teacher's part.

Other possible uses for AI include tutoring, assessing small group collaboration, and assisting with content for lesson planning. For example, AI could provide assistance in student group discussions in classrooms by providing insights to the teacher on students' collaboration dynamics or by directly facilitating group work — encouraging a constructive, equitable dialog and discouraging behavior that marginalizes certain students.

At the same time, participants expressed concern about using NLP to directly interact with students as AI tutors, without a teacher in the loop. Even if technical gaps were addressed to ensure that the model practices high quality instruction, AI lacks the uniquely human connection with students that can promote learning success. Thus, many of us envision applications of AI tutors that are mediated through or scaffolded by

a human tutor or teacher. For example, AI can enhance the human-student tutoring interaction by providing insights to tutors about students' understanding or generate response suggestions that tutors can accept or reject — this could happen real time in text-based contexts and as a post-session reflection tool in face-to-face contexts. Another promising example of AI tutors is in higher education, where such tutors can help respond to students in large introductory courses, since students often ask the same questions.

Some participants argued that in certain situations an AI tutor may be effective. For example, students can answer questions and learn from errors without a fear of judgment or stressful social interactions. Furthermore, in the absence of human tutors, AI may still be better than not having a tutor at all. Such an AI tutor can provide feedback and explanations, helping resolve learner misconceptions in real-time without giving answers, a critical path to math attainment especially.

Lastly, teachers can use AI to locate useful content and suggestions for their curriculums, including videos, summaries, and alternative explanations for students. These content suggestions could be used to personalize instruction for diverse students, or for students who are multilingual.

All these potential applications serve one purpose: to help teachers be more effective with the time that they have. Whether teachers use NLP technology to uplevel their own learning, provide more personalized support to students, or orchestrate cohort activities to increase engagement, it is likely to become a daily aid within and beyond the classroom.

Begin with Equity. When teachers have more time, they are able to focus on students that need more support. This could have a deep impact on students in large and under-resourced classrooms.

Centering Teacher Needs. AI tools could help mitigate burnout by reducing time spent on tasks that teachers experience as busywork, both during the school day and in the evening, when many teachers work beyond their contracted hours.

***Promote High-Quality Instruction*.** With increased time and new insights, teachers will be able to focus on accomplishing their instructional goals.

Build and Inform Educational Theory. By comparing different approaches to performing teaching tasks (human only, AI only and blended approaches), we can better understand the unique contributions of a human teacher vs an automated system in facilitating student learning.

Enhancing Professional Learning

Teachers receive feedback inconsistently, and they rarely receive low stakes, non-evaluative feedback. Yet causal evidence suggests that such feedback is a key lever to improving instruction and student outcomes.[7] NLP is already being used to provide teachers with on-demand automated feedback, which teachers can view privately or with the assistance of a peer or coach. Amplifying and extending this work can increase teacher access to feedback and has the potential to improve classroom instruction.

An automated feedback cycle begins when teachers audiorecord a lesson and upload it to a platform that analyzes the teacher and student talk in the lesson. For example, the TeachFX platform provides reports on a lesson's student-teacher talk ratio. M-Powering Teachers is a tool that provides in-depth analysis of teachers' uptake of student ideas. Similar reports could also be made for student reasoning, disciplinary practices, and the use of scientific or mathematical language, among other metrics

[7] Emily Boudreau, "The benefits of low-stakes teacher evaluation," Harvard Graduate School of Education, November 2, 2019, https://www.gse.harvard.edu/ideas/usable-knowledge/19/11/benefits-low-stakes-teacher-evaluation.

inside the classroom. In all examples, teachers view and interpret their data, then use it to think about ways to improve their next lesson. Recording the next lesson allows them to see progress toward their goals.

Achieving the promise of NLP-based automated teacher feedback relies on making progress in three areas. First, automated speech recognition — systems used to transcribe teacher and student talk for analysis — must be adapted for noisy classroom environments, and for the kinds of speech patterns and language used in K-12 classrooms. Second, we need a range of measures. Some of these measures will focus on content-specific classroom processes outside of mathematics, where most early work has occurred. Other measures will capture hallmarks of equitable classrooms, including inclusive teacher instructional moves, teacher and students' use of unbiased language, and equitable student participation.

Working with teachers themselves, we need to understand how best to create automated feedback delivery systems that teachers find both appealing and useful. Such systems may include not only the automated feedback itself, but embed that feedback in coaching routines that can be carried out by local or virtual coaches. AI-based coaching — with discussions supported by chatbots — can help teachers set goals and work toward achieving them.

Begin with Equity. Teachers are not always given the tools and support to help students with every learning need. Data and targeted professional learning will help teachers give support to all students. Furthermore, metrics on equitable interactions can inform teachers about subtle inequities in their classroom that they may not be aware of.

Center Teacher Needs. Professional learning tools and automated feedback can provide teachers recognition for making improvements in their classrooms, immediately and positively reinforcing progress.

Promote High-Quality Instruction. Automated feedback can focus on improving teaching methods known to increase student learning, such as academically productive talk moves.

Build and Inform Educational Theory. The data collected in classrooms will inform our understanding of teachers as both learners and educators. We will both be able to better understand what classroom interventions are effective for students and what professional learning strategies are "sticky" for teachers.

Facilitating Adaptive Lesson Planning

When creating lesson plans, teachers adapt, modify and apply curricula adopted by their districts to their unique classroom contexts. AI can assist teachers confronting sometimes overwhelming choices by analyzing classroom data and making personalized recommendations. According to a March 2023 survey conducted by the Walton Family Foundation, just three months after the ChatGPT was released, 40% of U.S. teachers were already using the software, with their primary use case being lesson planning. This number jumped to 60% by July, despite ChatGPT's observed flaws, such as its hallucinations and inability to represent a diverse range of voices. Could we maximize the benefits of language technology to save teachers time and improve the quality of their lessons?

We already see that AI can be helpful in generating ideas for activities and worksheets, but it is unclear the extent to which AI tools are effective at adapting curricula to learners' individual needs. Working with teachers and students to enhance large language models (LLMs) — e.g., by crafting prompts and fine-tuning data, so that

the models can effectively adapt learning materials — shows great promise for improving students' experiences and learning. For example, Individualized Education Plans (IEPs) are created after careful discussion with and observations of the learner. NLP technology can help locate and synthesize critical utterances, concepts, or issues that a learner expresses, and use these insights to suggest an IEP for the learner, which a specialist can then review and edit.

Furthermore, these tools can be especially effective when used for subjects that do not typically have packaged curriculum materials, such as secondary ESL, or for helping students catch up on prerequisite material.

Begin with Equity. Using off-the-shelf LLMs (e.g. ChatGPT) without careful prompting and tuning is unlikely to produce culturally responsive teaching materials or materials that individually cater to students with special learning needs. We need to adapt and carefully validate any LLM-based lesson planning approach to ensure that they are successful at personalizing materials to students who are marginalized or those with special needs. Furthermore, since LLMs are likely to learn from existing curricula, many of which have their own biases, we need first ensure equitable representation in data on which the models are trained or tuned.

Center Teacher Needs. Lesson planning interventions will complement teachers' expertise and reduce their workload required to create customized, high-quality lesson plans for students.

Promote High-Quality Instruction. These tools give students individualized curriculum attention they would not otherwise receive.

Build and Inform Educational Theory. These lesson planning tools and the data collected from them will give us insight on a granular level about how students learn and what curriculum interventions are the most effective.

Enriching Formative Assessment

NLP can change understandings of and responses to student performance. Typical assessments take periodic, distal snapshots into student learning and then require human analysis for interpretation on how to adjust instruction accordingly. While standardized testing can be an effective diagnostic tool and serves other purposes such as school- and district-level accountability, it also suffers from a variety of measurement issues, and frequent testing can be costly and counterproductive. NLP has the potential for enabling proximal, easy assessment by capturing students' learning in situ and analyzing student discourse in the classroom. In addition, it could enrich the diagnostic information provided through standardized testing by analyzing item-level student responses, especially for open-ended questions.

NLP can also augment assessments by facilitating the creation of personalized, adaptive, formative test questions. Such test items could increase accessibility for diverse populations and increase alignment with students' lived experiences.

Begin with Equity. Student achievement can be assessed with more equitable metrics. Potential bias in the models' assessments need to be measured and mitigated.

Center Teacher Needs. Automated grading could help teachers pinpoint where and why a student struggled, categorize their errors, and suggest next steps.

Deliver High-Quality Instruction. Teachers can use feedback from automated grading to gain new insights into students' strengths and challenges, guiding their instruction.

Build and Inform Educational Theory. These approaches could inform new models of educational assessment.

Challenges

There are several challenges that impede our ability to apply language technology in education in an equitable, safe and effective way — many of these challenges are related to the complex interplay of community members, technology, and research involved. In exploring these challenges, we chose to focus on those that could slow the development of high-quality datasets, obstruct the process of building effective classroom tools, and undercut the cultivation of research–practice-edtech partnerships.

Creating High-Quality Datasets

To create the most effective language technologies, we must first develop high quality datasets that enable training and tool-building for multiple purposes. New datasets, for instance, can enable the development of more precise measures of classroom variables such as equitable student talk, student-to-student deliberation, and the coordination of background texts and images with student and teacher talk.

Though student formal learning performance is an important metric, datasets may capture other influencing aspects of a student's life including belonging, confidence, and engagement. Research analyses can benefit from looking at triangulated data about students' experiences, including longitudinal data, basic metadata on teacher background, links between kids' utterances, learning management system (LMS) data, digital resource data, pictures of tasks students engage with, lesson plans, and "exit ticket" surveys that students and teachers take post-lesson about their experiences. Rigorous analysis of how these data, factors, and interventions interact is both possible and also required to understand the broader context of the classroom. At the same time, there may be a potential tradeoff between the depth and breadth of data being collected, and the right balance is dependent on the specific research project.

When capturing experiences in the classroom, we also have to ask, what are we missing? Our understanding of the classroom environment is incomplete without a window into nonverbal interactions. Video analysis can provide such a window, but audiovisual technology is currently out of reach for most classrooms. We would also like to understand interactions with school counselors and informal comments from school personnel that can have a huge impact on a student. That said, when collecting data in the classroom, something is better than nothing.

To collect data effectively, we must address these technical challenges:

- *Ensuring Equity:* Youth of color and historically under-resourced communities are often underrepresented in model-training data. This underrepresentation can perpetuate model biases.

- *Obtaining Accurate Transcripts.* Accurate datasets require accurate transcripts; we would aim to capture 100% of student talk in the classroom, processed by accurate speech recognition models. Transcripts should capture linguistic variation (e.g., use of different languages and dialects in the classroom) accurately, as well as all participation structures (e.g., whole classroom and group work).

- *Achieving Diverse Sampling.* In order to best help teachers and learners across backgrounds, data collection must be done with a diverse and representative sample of teachers. However, we recognize that there is no perfect definition of diversity for data and recognize that sometimes data collection can expand after it begins.

- *Maintaining Privacy.* Privacy is a clear priority. We must safeguard student data by storing it on encrypted servers and ensuring that only research team members with the right permissions have data access. Even if the researchers have permissions to access personally identifiable information (PII), they should use de-identified data for analyses whenever possible. Furthermore, teachers must be guaranteed that the data collected from their classrooms will not be weaponized to threaten them or their schools.

- *Collecting Unobtrusively.* Data must be collected in a way that doesn't interrupt or interfere with the classroom. This is challenging since it may require special equipment (e.g. one that can be hidden or very easily operated), additional human resources (e.g. someone to operate the equipment), all of which come with their own logistical and privacy-related issues.

- *Linking Language Data to Administrative Data.* Most of the time, language data and administrative data is obtained through entirely separate processes (e.g. classroom observation vs district databases). Linking such data is often important, and can be highly challenging. For example, developing measures of equity in classroom discourse requires mapping student talk to demographic information. However, doing so is challenging both logistically and in terms of ensuring student privacy throughout the data collection process. Even if the ultimate goal is a dataset with no personally identifiable data (e.g. de-identified transcript where speakers are tagged with relevant demographic information), collecting the data does require temporary access to personally identifiable information about students.

- *Scaling.* Developing the best approaches for language technology in the classroom requires robust evidence, which in turn often requires large, representative datasets. Collecting such data is challenging due to obstacles to teacher and district buy-in and heterogeneity of implementation (i.e. different districts may supply different types of data that may be hard to synthesize).

Recommendations

★ Teachers should gain from any data collection in their classroom. The data may give insights into their own practice. Teachers should have opportunities to work collaboratively to design NLP software, and to participate in the research process.

★ Ensure strong protections for students and teachers. When getting consent, researchers must clearly communicate how the data will be used for current and future research. They must specify who will have access to the data now and in the future.

★ Maximize the value and quality of newly collected data. Data collection can be resource-intensive and logistically complicated. Researchers should ensure that their data will meet criteria for answering a range of research questions before beginning the collection process.

★ Rigorously review and document the context of training datasets. Nonrepresentative training data can make tools ineffective and even exacerbate equity issues. For example, NLP insights gleaned from students using a discussion-based curriculum may not be valid or useful for districts and schools that use direct instruction.

★ Create different types of datasets. Some datasets should be small and intensive, following a smaller sample over a large period of time. Others must be broad, following a larger set with less intensive data collection.

★ States and local governments should invest in robust local, school district–level data infrastructure for collecting, storing, and managing these new datasets.

★ Funders should facilitate ethical, robust, and safe data-sharing infrastructure pipelines between school systems, companies, and researchers.

★ Funders should support research and development around data collection, in addition to the tools themselves. Testing different types of data collection equipment and improving fundamental technologies such as speech recognition can significantly enhance the quality of the collected data and the impact of the research.

Building Effective Tools for Classrooms

Often, technology and professional development are “one more thing” for teachers to navigate in their environment. Developers of new language technologies have the opportunity to build, from the ground up, tools that teachers find useful and effective.

Those aspiring to create these tools must start by asking the question: what do teachers want? Tools should be based on the teacher's true needs, not imagined ones. To design genuinely useful tools, developers must collaborate with teachers from the very beginning.

Once specific needs are targeted, the tools must be designed to be as easy to use as possible. User-friendly designs are crucial for the onboarding process; no teacher will want to integrate a tool that takes significant effort to learn or that proves to be unreliable. Also, any tools introduced in the classroom should combine well with other district systems.

Furthermore, teachers need to perceive real benefits to integrating these tools into the classroom, such as insights, saved time, or the ability to demonstrate progress on accountability measures (without risking penalties). Making tools mandatory could increase teacher participation, but it could also add to the already large pile of work teachers have to do, threaten their sense of agency, increase teacher burnout, and cause teachers to resent the tool.

Technical problems around collecting data have to be solved, otherwise “everything downstream suffers,” said Shyamoli Sanghi, machine learning engineer at TeachFX. Biased data and designs will only perpetuate more bias and related harm. Tools will be contingent on the data collection process, and the challenges of collecting high-quality datasets also extend to creating effective tools.

Open Questions

- How do we align agency, user desires, and outcomes? Customizable tools promote more agency and personalization for a classroom, but the extra complication can make onboarding much more difficult.

- How do we encourage adoption of the tool? We need to ensure that tools are trustworthy and empowering for both teachers and students. Tools should not undermine the authority of teachers in the classroom.

- Who should have access to classroom data generated through technology use? The data may be valuable to district officials, for example, but using it for teacher evaluation could have negative consequences on teacher well-being and lead to burnout. If the data is shared among colleagues, does it encourage teachers to share insights with each other about their classrooms or will it cause unnecessary competition and professional isolation? Competition could be a positive motivator but might also promote reporting loopholes "equivalent to erasing bubbles [on standardized tests] and filling them in." (Jim Malamut, Stanford University PhD Student)

Recommendations

★ Validate equity at all stages of the research and development process. It is crucial to ensure that tool implementations do not propagate inequities and biases.

★ Co-develop tools with teachers to make sure the tools fit teachers' needs and build on their expertise. Just showing a finished product to teachers is insufficient.

★ Ensure that tools provide teachers with mutually reinforcing, rather than conflicting information. For example, suggestions provided through automated feedback should align with principles set forth by the teachers' district or school.

★ Do not "move fast and break things." Unreliable tools could harm teachers or students and will diminish teachers' trust in new implementations.

★ Reduce the burden on teachers. Tools that feel like one more task will not be adopted effectively. Successful tools will increase teachers' enjoyment in teaching and improve the quality of their instruction. The gain from the tool has to be worth the cost of setting up the technology, orienting kids to use it, and interpreting results.

★ Onboard effectively. Left to themselves to explore how a tool works, teachers may have difficulty learning its full functionality, especially given their normal

workloads. Tools must be taught well. For example, developers could work with instructional coaches to understand how they might facilitate teachers' use of automated feedback.

- ★ Do not make assumptions about classroom resources. We believe it is safe to assume that teachers will have phones, but tools that require more technological resources are not realistic in an average classroom scenario.

- ★ Funders should incentivize partnership between developers, teachers, and system leaders.

- ★ Funders should invest in solving fundamental technological challenges, since those are critical to building effective tools in classrooms.

Cultivating Research–Edtech Partnerships

We have good reason to be skeptical of edtech. Most products marketed to schools, districts, and families lack evidence of effectiveness; and no organization provides oversight of quality. There is no FDA for edtech. A recent report by EdTech Impact, a U.K.-based independent review platform, found that just 7 percent of edtech companies used rigorous evidence of impact.[8] Instead, the success or failure of edtech usually depends on the skills of marketing and distribution teams. Lack of engagement with evidence isn't confined solely to edtech companies and developers; it applies to the entire edtech ecosystem. Many buyers of edtech products and services do not look for robust evidence showcasing effectiveness. A national survey of 515 school and district leaders responsible for edtech procurement decisions, organized by a working group at the Edtech Efficacy Research Academic Symposium, showed that only 11 percent request peer-reviewed research.

Edtech is, almost by definition, continuously changing and improving. As a result, we cannot expect every program, and certainly not every iteration of a program, to have high-quality evidence of its effectiveness on long-run outcomes. That research takes time. Nonetheless, educational institutions and other edtech consumers are better served by programs that build on research findings, employing approaches that have been shown to work for learners. Outside of edtech, EdReports reviews the quality of instructional materials and assesses them based not only on direct randomized

[8] "Edtech Should Be More Evidence-Driven—EdSurge News," *EdSurge*, June 3, 2022, https://www.edsurge.com/news/2022-06-03-edtech-should-be-more-evidence-driven.

controlled trials of their effectiveness, but also on whether they leverage approaches based on learning science, which makes them more likely to be effective.

Edtech could yield more impactful results if it were both to engage more fully with ongoing research and to build on the currently existing research knowledge about teaching, learning, and engagement. Learning scientists often promote iterative design and implementation of digital tools in learning environments using a research methodology called design-based research, or DBR (Design-Based Research Collective).[9] DBR is a research paradigm that seeks to design, implement, evaluate, and iterate upon educational improvements and solutions via testing in real-world contexts.

To ensure that new technological innovations are rooted in core problems of practice, researchers and practitioners often form research-practice partnerships (RPPs), which are long-term collaborations among researchers and relevant education partners that aim to promote systemic educational change by producing and using research evidence related to matters of shared concerns and aims. RPPs are intentionally organized in ways that attend to the values of community members, power, and history of local settings.

More recently, this model has been expanded to capitalize on the burgeoning market of edtech solutions while integrating educational research in the design of new tools for learning. These new **research-practice-*industry* partnerships (RPIP)** represent a co-design method of research and development characterized by partners representing diverse perspectives working to solve a common problem. RPIPs involve methodological adaptation by researchers to broker tool implementation and translate user feedback quickly, and industry commitment to tailor solutions to the needs of practitioners.

Partners	Value They Contribute	Key Challenges
Researchers	Researchers have potential to advance research methods and theory to better understand how technology shapes learning. Researchers help develop the theories,	Ethics: Conflicts of interest are possible, and researchers must remain independent. Perception/Policies: Researchers may be seen as “sellouts” if collaborating with industry, potentially negatively impacting their rate of publications. Publications/Timescale: Research often takes years; reputable research journal publications

[9] Design-Based Research Collective. "Design-based research: An emerging paradigm for educational inquiry." *Educational researcher* 32, no. 1 (2003): 5-8.

	processes and structures on which the products are based.	take 6-24 months for review, and an additional 12 months to publish. How do we build towards something greater that is both valued and expedited?
Practitioners	Practitioners have a lot of expertise, feedback and ideas that must be heard, especially about what tech does and doesn't work and how research results could be useful.	Time: Given the time required for rigorous research and data collection to take place, RPIPs may add to teachers' already high workload. Onboarding: New users take time to become familiar enough with a tool to deploy it in the classroom. Infrastructure and Resources for Engaging: Teachers, school leaders and district leaders lack funding, incentives and internal capacity for engaging in research, development of data infrastructure, and routines for working with research and industry partners.
Industry Partners	In RPIPs, industry partners often have the most resources and infrastructure to translate input from practitioners and researchers into scalable, positive impact.	Organizational Structure: Industry partners must thoughtfully structure their business to include all stakeholders of RPIPs. Return on Investment (ROI): RPIP work may not have an immediate ROI and require substantial time and upfront costs that can be difficult to prioritize for companies. Longer-term, RPIPs may better fit the needs of small and midsize enterprises (SMEs) than large companies. It may seem more efficacious and efficient to create research teams/market research teams internal to the company. What is the long-term importance and impact of external partnerships? Growth Mindset: Inconvenient truths about the quality of the product can frequently be challenging to hear, requiring a mindset orientation toward continuous improvement.

Open Questions

- Who is the matchmaker between researchers, practitioners, and industry partners? How are those relationships managed? For a product to work effectively, it should be matched to a district that can make use of it — where it corresponds to students' and teachers' current priorities and needs.

- How can we ensure equity while using our networks to build partnerships? Finding partners is a trust-based practice; a partner is likely to call on the same voices they have worked with in the past, consequently overlooking historically unengaged communities, schools, and teachers. Furthermore, partners are likely to work with other partners with the bandwidth and permission to participate, which can skew towards white, male, and resource-rich partners.

- What are the key methods to drive RPIPs beyond those shared with design-based integrative research?

- How do we align timelines for each party? Industry partners tend to have an accelerated view of the world; frequently they are trying to make the most happen in the shortest time possible. Practitioners, on the other hand, usually look ahead one or two years. Researchers tend to think in a wider view, taking time to collect, analyze, and write up research that typically spans multiple years.

- How do we manage intellectual property (IP) concerns? IP worries for individuals, companies, and universities could shut down otherwise productive partnerships. Industry partners need to stay in business, but academics and practitioners should not surrender key IP to organizations in this process. We need a solution that allows innovation and unites different stakeholders.

- Who will invest in partnerships between researchers, ed tech industry leaders and practitioners? What incentives exist to support these partnerships?

Recommendations

- ★ Engage all partners at the conceptualization phase and create routines for engaging in each subsequent R&D phase. Product designs must be informed by research and theories of learning, as well as the priorities and day-to-day experiences of school systems. Within industry, we see success when researchers are embedded early in companies. That way, even if these companies turn to academia for supplementary research, they have point people to engage with every step of the process. Examples include Age of Learning, Amira Learning, Imagine Worldwide, and MainStay, all of whom have conducted extensive research on their products and built early internal research teams.

- ★ Listen to and co-design with practitioners. Teachers and school system leaders have a lot of feedback and ideas about technology that works and doesn't work. We need to ensure their needs are accommodated and that they know they are heard.

- ★ Researchers should prepare for quick turn-arounds of analyses. When working with designers, they can learn from partners how to innovate quickly and think creatively about publications and academic products. In the academic space, researchers need to speed up the research process and publication cycles. Education researchers could explore more expedient peer-review models (e.g., Association of Computing Machinery (ACM)) to improve educational research reporting. In addition to helping align timelines, this would help research get published faster, with high-quality peer review, and generate more citations through open access publication. This could also mitigate concerns about potential disputes over IP control.

- ★ Researchers must take a true third-party objective stance to partnership. It can be uncomfortable to tell partners that they are not ready to deploy or a major program or investment is not "working," but when that is the case, the point should be made.

- ★ Distinguish whether you are engaging with practitioners as a partner or participant. When partnering, it is crucial that we show that we value teacher and school system leader expertise, instead of prescribing without listening to the problem. We need to emphasize "engagement" over "recruitment," fostering partnerships that benefit teachers and leaders without requiring huge sacrifices from them. As part of building trust, we recommend sharing data with teachers and leaders so they can see the positive impact that they are making.

- ★ Provide practical reasons for teachers and school system leaders to engage. For example, fellowships like Hollyhock at Stanford University, that welcomes high school teachers from across the country for two years who are interested in deepening and developing their content-specific instructional practices, can enable teachers to be part of the development process for different tools and give them the skills they need to use them. System-level research-practice-industry partnerships provide resources and infrastructure that encourage school and district leaders as well as teachers to engage in research and development endeavors. These initiatives would not be possible without financial support.

- ★ Researchers should provide thorough literature reviews for their partners in accessible formats and venues (e.g., not only in peer-reviewed journals with paywalls). Shared readings, terms, and prior findings can be used to ground designs.

- ★ Funders should consider a variety of funding models, including seed grants and fellowships (e.g., academic fellowships for edtech professionals), in order to foster RPIPs. In recent years, funders, ranging from the government to foundations, have supported new funding models to bridge research and industry, and accelerate the diffusion of research-anchored solutions. For example, federal efforts have included new pools of funding from IES (ARPA-ED) and NSF (Convergence Accelerator, SBIR). Foundations and universities, such as the Gates Foundation and the Stanford Accelerator for Learning, offer seed grants and an accelerator studio to support research-anchored solutions developed by edtech and educators in partnership. The William T. Grant Foundation and the Spencer Foundation have grant-making programs that develop infrastructure, institutionalization, and transformation for partnership work, helping to change universities and their practice partners.

Conclusion

This conference and whitepaper are intended to open up further conversations. As researchers and edtech entrepreneurs explore avenues to incorporate language technology in the classroom, we must continue expanding our understanding of its uses and potential, while hewing to a positive vision: we want to create innovative tools that are born from the shared goal of equity across classrooms and that will uplift underserved students and teachers.

A positive vision is imperative, but not on its own enough. Iterative development and rigorous evaluation will be key in creating technologies that can help realize the vision of equitable classrooms with high-quality instruction. This can only be accomplished through intentional collaboration across areas of expertise and by involving practitioners from the outset, following the “AI in the loop, educators in charge” principle while eschewing the “move fast and break things” approach.

Our hope for the future of NLP technologies is linked to our excitement about a field that can also form a community, one with shared values and a common mission: promote equitable education; develop accessible, research-informed tools; and, by using those tools, learn more about human learning.