How Can We Co-Design Learning Analytics for Game-Based Assessment: ENA Analysis

Yoon Jeon Kim [0000-0002-3467-1238], Jennifer Scianna [0000-0003-1029-3452], Mariah A. Knowles [0000-0002-1798-4830]

University of Wisconsin – Madison, Madison WI, USA

yj.kim@wisc.edu, jscianna@wisc.edu, mariah.knowles@wisc.edu

Abstract. The broader education research community has adopted co-design, or participatory design, as a method to increase adoption of innovations in classrooms and to support professional learning of teachers. However, it can be challenging, due to co-design's dynamic nature, to closely investigate how the co-process played out over time, and how it led to changes in teachers' perceptions, beliefs, and/or practices. Applying Quantitative Ethnography, we investigate how teachers and researchers collaboratively designed assessment metrics and data visualizations for an educational math game; we discuss the interactions among the co-design activities, teachers' learning, and qualities of the dashboard created as the output of the process.

Keywords: co-design, ENA, teacher professional learning, data visualization, human-centered learning analytics, game-based assessment

1 Introduction

By implementing educational games in classrooms, teachers can provide authentic and engaging opportunities to support learning of academic content as well as valuable cognitive and non-cognitive skills [1]. The advocates of game-based learning long recognized affordances of data generated from gameplay for improving teaching and learning in classrooms. For example, game data can be processed and presented to provide teachers with greater insights into students' learning, so they can provide timely feedback [2]. Making these data actionable and meaningful to teachers, however, poses several challenges. First, teachers must understand what kinds of data (or evidence) were collected and processed related to which learning outcomes. Second, teachers must be able to make sense of the presented data and trust its accuracy and validity. Third, data visualization tools coupled with games must be usable by teachers in real classroom contexts. In summary, making game-based assessment data useful to support teachers' pedagogical decision-making in classrooms is not trivial, and the disconnect between teachers' needs and learning analytics development has been consistently

discussed as one of the main barriers to fully leveraging the data affordances of educational games, and educational technology more broadly [3].

In this paper, we aim to investigate the interactions among teachers' perceptions about assessment, data visualization and learning analytics, and the co-design process in the context of game-based assessment. We investigate these interactions in an iterative development process that engaged teachers as co-designers to develop teacher-facing, interactive dashboards for Shadowspect– a 3D puzzle game for assessing Common Core Geometry standards, student persistence, and spatial reasoning.

2 Theory and Relevant Work

2.1 Co-Design

Co-design, participatory design, and co-creation all have the goal of involving stakeholders as collaborative designers [4]. In education research, co-design is increasingly adopted as a form of design-based research (DBR) that incorporates several types of stakeholders in the development and research processes. In this view of co-design, designers (or researchers) and stakeholders are on equal footing in the design process albeit with diverse roles. Teachers become designers seen as "experts of their experience," while designers become facilitators easing teachers' expression of creativity and as product experts under design.

As illustrated in Figure 1 (adopted from [5]), a typical design process follows four phases: pre-design, generative, evaluative, and post-design. The first dot in the process indicates the determination of the design opportunity (or defining the problem), and the second dot indicates the finished product. While traditional design process brings the users in at the back end of the process, co-design aims to get end users involved in the front end. The key ingredient of successful co-design centers on an iterative and creative process of "making things" that illustrate future opportunities, concerns, values, and views on future ways of doing or living.

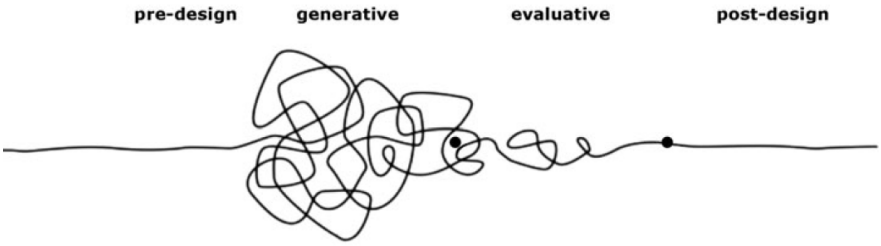

Figure 1. Illustration of a design process

The co-design research community has been accumulating various co-design methods and tools to meaningfully engage end users who are not trained as professional designers in this creative process. Sanders and Stappers [5] describe three types of activities that occur during this creative, making process: probes, generative toolkits, and

prototypes. The probes intend to get users to reflect on and express their experiences, feeling, and attitudes in forms and format that provide inspiration for designers. The generative toolkits intend to engage users to make expressive artifacts—artifacts that might not be directly related to the actual end product but demonstrate how they perceive and envision future opportunities and discuss them. These expressive artifacts can then be reviewed by the designers to identify underlying patterns and themes that can inform the prototype. Some better known toolkits include User Persona, User Journey Map, Role-Play, and Photo Studies (for the whole slew of toolkits see, for example, Kumar [6]). Prototypes are physical manifestations of ideas and concepts with varying degrees of fidelity and usually come later, closer to the evaluative phase, to get feedback from the end user. In summary, the key element of the co-design process is an iterative back and forth between divergent and convergent activities that are facilitated by the designers (in our case, design-based researchers) through "creative acts of making", in which both designers and users actively participate.

Beyond the immediate benefits of leading to products that better align with the end users' needs and values, many researchers have begun acknowledging the additional benefits of having teachers participating in this creative process as teachers' professional learning opportunities [7, 8]. For example, Voogt et al. [8] report that by engaging teachers as designers, teachers could learn how an innovation works rather than simply that it works.

2.2 Teachers' Assessment Literacy for Technology-Enhanced Learning Environments

Currently, one of the main challenges to developing meaningful learning analytics and complex assessment models for teachers is the limited understanding of what kinds of assessment (or data) literacy teachers need to have to fully leverage assessment in technology-enhanced learning environments [9]. The general tone is that teachers lack the requisite skills and expertise to make full use of the data available to them from interactive educational technologies [10].

One well-documented challenge for teachers' use of highly processed machine-learning driven assessment models is the non-transparent and inscrutable nature of these algorithms [11]. For example, if a teacher receives an alert message that a student is 90% likely to quit, the teacher might want to know both why the student is becoming disengaged and how the 90% estimation was reached. In most cases, however, no support is provided; teachers must interpret such estimations on their own. Additionally, as teachers may not necessarily be fluent with machine learning concepts, they may also struggle to critically examine these algorithms. In situations where teachers are asked to simply trust the outputs without understanding their intricacies, the "black-box" nature of the algorithms may lead to mistrust or uncertainty in the models and their results. Although recent work has begun to improve the interpretability of such algorithms, less attention has been paid to understanding what skills and knowledge teachers require and what features and qualities developers should consider to support teachers' use of such technologies.

In the current project, we define a teacher with assessment literacy in the context of educational games as follows [12]: (1) value non-academic, nontraditional, and process-oriented skills and attributes of learners that game environments can support; (2) understand what these constructs mean and can identify possible evidence for those constructs based on students' gameplay; (3) critically and curiously investigate how the data was processed, based on what rules, and understand the role of computing and artificial intelligence and its limitations even if it is not fully understand how the algorithms are being built; (4) use data and visualization tools to identify strengths, weaknesses, growth, and productive and unproductive struggles of learners beyond proficiency; and (5) strive to gain new, delightfully surprising insights about learners that they couldn't see with traditional forms of assessment; and finally (6) explore and dig into the data at various levels (i.e. individual, subgroup, classroom, grade) and with diverse goals (e.g. what's the puzzle that everybody is struggling with, so I can intervene?). This describes the secondary, professional development goals of the design team for the teacher fellows in addition to the creation of a meaningful, serviceable analytics platform.

3 Context

The research team selected 8 math teachers as *design fellows* from 16 secondary school teachers who applied in response to an open call for participation. The teachers were selected based on their interests in the educational value of games, data use in their classrooms, and interest and prior engagement in co-designing processes. The team and teachers met monthly during development iteration cycles for 12 months. A typical co-design session lasted 2 hours. Due to COVID-19, all design sessions were conducted and recorded remotely via Zoom. The team collected several sources of data: design session discussions, teacher interviews, teachers' individual think-alouds, artifacts generated by the fellows, and the team's field notes.

The focus of individual co-design activities varied from generated activities using digital and nondigital toolkits to evaluative activities using prototypes. For example, the teachers used a digital toolkit called *Caterpillar* to come up with different instances for how persistence would be demonstrated in student's gameplay and how they would interpret them (Figure 2a). The teachers also collaboratively created visualizations for possible prototypes (Figure 2b), and later evaluated how these prototypes worked or didn't offering ideas to improve them.

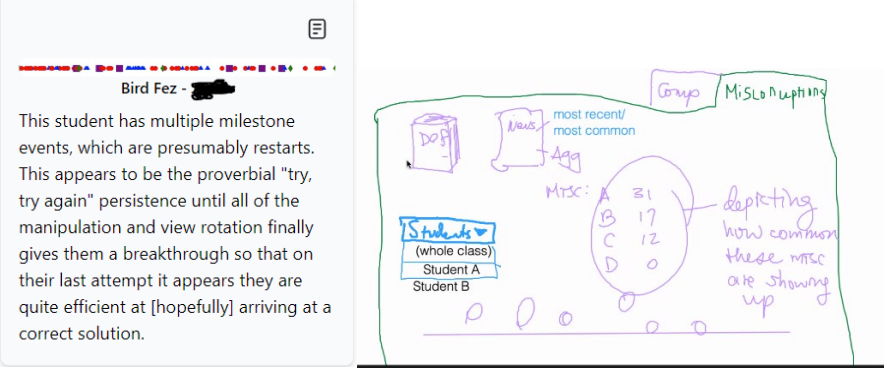

Figure 2a and 2b. Examples of co-design activities

For this paper, we analyzed all sessions, entry and exit interviews, and individual think-aloud activities to investigate the interplay between teachers' assessment literacy and the co-design process. To these ends, we address two questions:

RQ1: How did teacher discourse change throughout the co-design process to demonstrate changes in their thinking about the connection between assessment and student learning in game contexts?
RQ2: If change is demonstrated, how did the co-design activities support it?

4 Methods

4.1 Data and Pre-Processing

Transcripts were generated based on audio recording and checked for accuracy. Transcripts were assigned metadata to compartmentalize the activities speakers were partaking in during each session; a single co-design session may have included several activities, some of which were based on group work. Activities were then labeled as the type of activity they were designed to be: probes, toolkits, and prototype reflections. Turns of talk were further segmented into individual sentences to isolate concepts the teacher fellows were bringing to the forefront of their discussions.

4.2 Coding

The research team approached the data by conducting a thematic analysis which yielded commonalities between the ways teacher fellows discussed their use of dashboards, desires for teaching, and thinking around artificial intelligence. These themes were iteratively defined into codes that incorporated elements of the role of technology (*AI*), teacher actions with the technology (*Manipulate*), teacher affect (*Trust*), teacher goals (*Teaching*), and teacher understanding of students' performance (*Comparison, Performance, Sequence, Thinking*) (See Table 1 for code definitions and examples). The researchers used the web-based nCoder [13] to reach agreement on the identified codes.

Table 1. Code Book - Kappa and Rho scores from nCoder are reported in order of Rater 1 vs. Classifier, Rater 1 vs. Rater 2, and Rater 2 vs. Classifier. *rho < 0.05, **rho < 0.01.

Label	Definition	Kappa
AI	algorithms, intelligent systems, automation, and artificial intelligence terminology, eg. "The computer program…instantly corrects the work for the students and gives them that feedback as soon as they submit it at the end"	.94* .96* 1.0**
Comparison	comparisons between student performance and other points of comparison, such as other students, class averages, explicit standards, and implied expectations, eg. "I can compare my students, how they're doing within school, which is like our district and school"	.97* .96* .96*
Manipulate	combining, filtering, sifting, selecting, or otherwise shifting one's perspective of data in general, such as sifting through student work, separating students, or identifying students, eg. "Or sorting by, like, the most missed question"	.96* .96* .93*
Teaching	things that teachers do, ie., pedagogical actions and strategies for instruction and intervention, such as checking in with students and planning concept review lessons, eg. "I would wanna sit down with that student and help them to develop strategies"	.96* .92* .92*
Performance	quantitative measures of student achievement, eg. "And again, I mean, just, I don't even know what growth looks like, but there's gotta be a way to measure it"	1.0* 1.0* 1.0*
Sequence	statistics, features, and descriptions of how students choose to link together, order, skip, or repeat complete educational tasks such as a problem, level, or assignment, eg. "So like, you know, a badge for, like, coming back to a level that you originally skipped"	.92* .91 .91
Thinking	what actions, thinking, or lack of those things, students performed within an educational task, eg. "And the other thing that I noticed, too, was that Player 1 was the only one who, like, changed the perspective so you weren't viewing it on an angle"	.93* .97* .97*

Trust	validity of metrics and trustworthiness of algorithms, eg. “Reliable's not the right word, but, it makes me question, like, how valid these scores are”	1.0** 1.0** 1.0**

4.3 Epistemic Network Analysis

Three models were generated using Epistemic Network Analysis to identify different elements of the teacher fellow discourse throughout the co-design experience in line with the research questions. Units were identified as Speakers in a given Session. As the focus of this study is on the teacher fellows’ discourse, utterances by the design team were filtered out of the model.

Conversation size was defined as all lines belonging to a single group conversation during a particular activity in one session, e.g. three teachers in a breakout room using a toolkit activity to identify metrics of persistence would be one conversation. This segmentation allowed for comparison of the benefits of each activity type in the ENA plots. Furthermore, each model used the same moving window of 16 lines. This size was chosen to account for both longer teacher responses during think alouds and interviews (so the entire response would connect to itself) as well as rapid-fire communication during the brainstorming sessions where teachers were providing many, short responses to one another. The primary difference between the three models were the rotations used for visualization.

To address RQ1, we created a trajectory plot [14] (Model 1) to show the distribution of points “through time”. In Model 1, the axes were rotated to maximize the variance between all sessions. This allows for better visualization of which codes and connections were dominant in the beginning, middle and end of the co-design process. To further address RQ1 about teacher fellows’ growth, we created Model 2 using a Means Rotation to separate entry and exit interviews with the design fellows, which were the only two activities included in the model.

RQ2 considers how the co-design activities engaged teachers in discussion of student performance, metrics, and data literacy. Model 3 used hierarchical epistemic network analysis [15] to compare teacher discourse when they were interacting with prototypes, toolkits, and probes. We selected hierarchical epistemic network analysis because it allows us to see the independent effects of two binary variables simultaneously. The x axis in this rotation was defined by the presence of Probes while the y axis was defined by the presence of Toolkits. The third quadrant, where both probes and toolkits are absent, is where prototype activities are identified.

5 Results

5.1 Co-Design trajectory

RQ 1 focuses on the movement of design fellows through the co-design process. Model 1, the MCR trajectory plot, demonstrates that teacher fellows began the codesign experience focusing on *Performance* and *AI* before moving towards *Thinking* and *Teaching*,

focusing on metrics of *Compare* and *Performance* while *Manipulating* data in the prototype before moving towards *Trust* and *AI* towards the end of the process (See Figure 3). The x-axis can thus best be described as moving from a Tool orientation on the left side of the plot to a Trust orientation on the right. The y-axis is best described by the nature of the conversations. Conversations that are connected to *Performance* tend to focus on the scores and metrics that students are getting both in the game under study and in school more broadly. Thus, we define the lower end of the y-axis as being centered on quantitative discussions. The positive end of the y-axis centers most on *Sequence* and *Thinking* discussions which are both descriptors of what students are doing. Thus, we label this end of the axis as being focused on more qualitative ways of thinking and discussing performance.

Model 1 had coregistration values of .96 and .94 along the x and y axis respectively. This demonstrates that the plot is an accurate representation of the data. Each axis was responsible for explaining 15% of the variance within the data.

Early in Session 2 of the co-design workshops, teachers contemplated if *AI* could even be useful for the problem they were hoping to solve regarding assessment:

> Teacher 1: But again like you talked about, the kind of technology that's out there with AI and your voice mail transcribing everything into texts, the possibilities, very soon, where all that stuff is going to be all set.
>
> Teacher 2: I wonder if there is some technology that could see what they're doing and identify common mistakes. [...] I guess a systematized way to measure progress would be great. [...] Can a computer put our students into those categories related to their process?

While teachers were familiar with technology, they were unsure if computers and *AI* would be the right tools to be able to evaluate student *Performance*. By the end of the workshop, teachers were thinking more deeply about the implications of bringing *AI* into the classroom: “So we need to think about like, it's a for-profit company that makes this stuff. So like who makes it? And what kind of biases do they have? Because if it's all like white men who come up with the algorithms and, so anyway, it makes me think about that.” Bringing ethical considerations into their discussion demonstrates a level of expertise with how algorithms and *AI* are made, not just that they exist within the dashboard program.

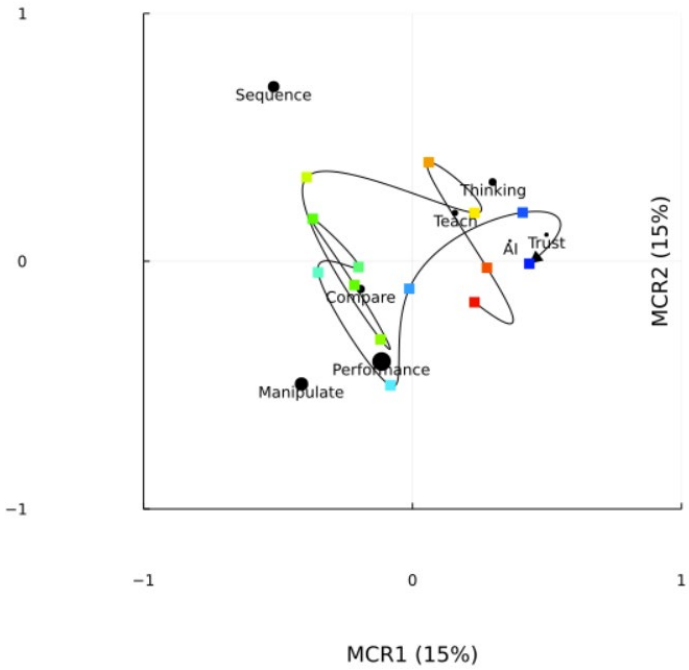

Fig. 3. The trajectory plot uses axes of MCR 1 and 2. Design fellows began and ended their experiences discussing *AI;* they began by focusing on *Performance* and *AI* and moved towards *Trust* and *AI* towards the end of the experience.

5.2 Teacher discourse shifts

Model 2 is visualized as an ENA plot (Figure 4) that includes only entry and exit interviews that were conducted 1:1 with the design fellows. The entry interviews show strong relationships among *Manipulate/Performance* and *Manipulate/Teach*. The strongest connection that emerged in the exit interview is the connection between *AI/Performance*. Model 2 demonstrates the shift in teacher discourse from discussing their use of dashboards and analytics tools in the pre interview to their understanding of dashboards as a computer-generated tool that they can control to show the nuances of student work and learning.

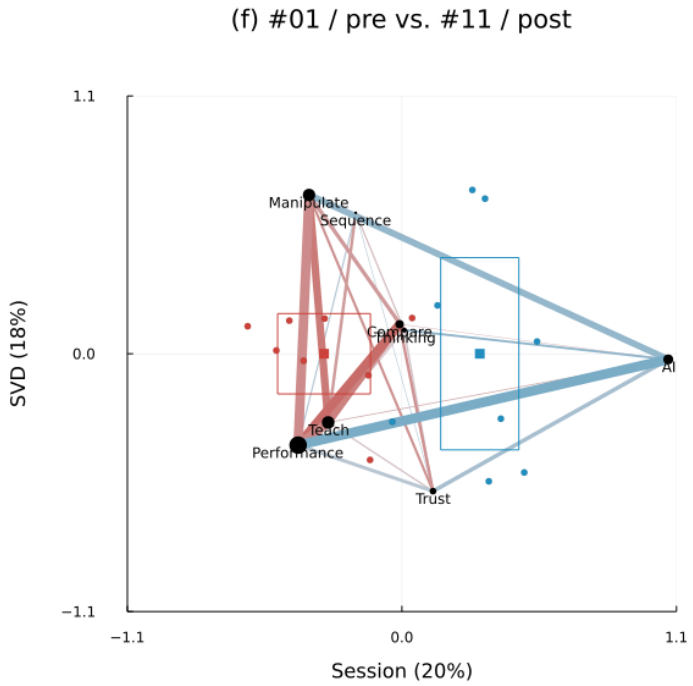

Fig. 4. A defining feature of the teacher post interviews is their focus on *AI.*

Interestingly, teachers were seemingly able to operationalize their *Trust* in *Thinking* and *Manipulation* during the pre-interview to be offloaded onto *AI* by the post interview. One teacher remark exemplifies this shift:

> "I had found something like we all did, when they put together teacher stuff, they try it, and students, and actually see how they feel about it, right? And so I think my trust issues go with like, like with the algorithm. Because I mean I don't fully understand the math or how it works, but I can trust that it's taking some important stuff that we teachers [inaudible] or that we value or weigh it as we think it is."

This teacher fellow is drawing the human sensemaking process into their interpretation of an algorithm's output. While they acknowledge that they do not feel capable of the math behind the *AI,* if the output comes from being designed and operationalized with teachers, the output can be meaningful.

Additionally, while teachers make a greater connection between *AI/Trust/Thinking*, they also bring in more about *Manipulation* and *Performance* in connection with *AI*. One teacher compared the impact an *AI* could have on their *Teaching*, especially with students they perceived as under-performing,

> "You have like the really high-performing kids that you know, and the ones who are way behind that you know. And then the rest of

> them are kind of just like lost in the shuffle. And so having like the AI kind of make a suggestion or be like, hey, watch out for this kid because lately we haven't been doing this. [...]Like that AI to be able to, and I don't know if I'm using that correctly, but that AI to like be able to keep track of all that stuff."

It seems that the activities the teacher fellow participated in allowed them to improve their connections to *AI* to include ways that they could operationalize it in their classroom.

5.2 The Value of Co-Design Activities

To tease apart the affordances of the different co-design activities, Model 3 used a hierarchical rotation to visualize the impact that probe and toolkit activities had on teacher discourse. When both probes and toolkits were present in the activities, there were strong connections between *Trust* and *Performance* as well as *Thinking* and *Sequence*. When probes were used without toolkits (noted with the blue down arrow), there were more likely to be connections between *Thinking* and *Teach*, and *Performance* and *Teach*.

Prototypes (shown as the red down arrow in Figure 5 where both toolkits and probes were absent) largely focus on the connections between *Manipulate* and *Teach*. Teachers were active in reflecting on the value of the prototypes as they were interacting with them. They often played out scenarios of how they could use the tools with minor improvements to augment the *Teaching* they could do in the classroom: "That's why it's almost like, if we can come up with, like, filters that we know give important information, like: these are the students you should go help, or these are the students who you should reward."

Model 3 depicts some of the trends noted in the prior two models. While Model 1 shows the teacher fellows moving towards *Teach* and then *Performance*, *Manipulate*, and *Compare*, Model 3 correlates those motions to teacher fellow participation in probes early in the co-design process and toolkits towards the middle. One fellow clearly connects *Sequence* to *Performance* as they are trying to discern student behavior during a sense-making toolkit activity: "And I think, this is the only one that I would, like, like to see in a different way, the reattempts after failure. Cause I'm like, "Compared to the class, you only tried to rotate it eight times. Like, saying students that got better scores is because they tried to rotate it more times." The teacher fellow was able to articulate why the metric they were able to play with was not adequate for understanding why students were performing certain actions.

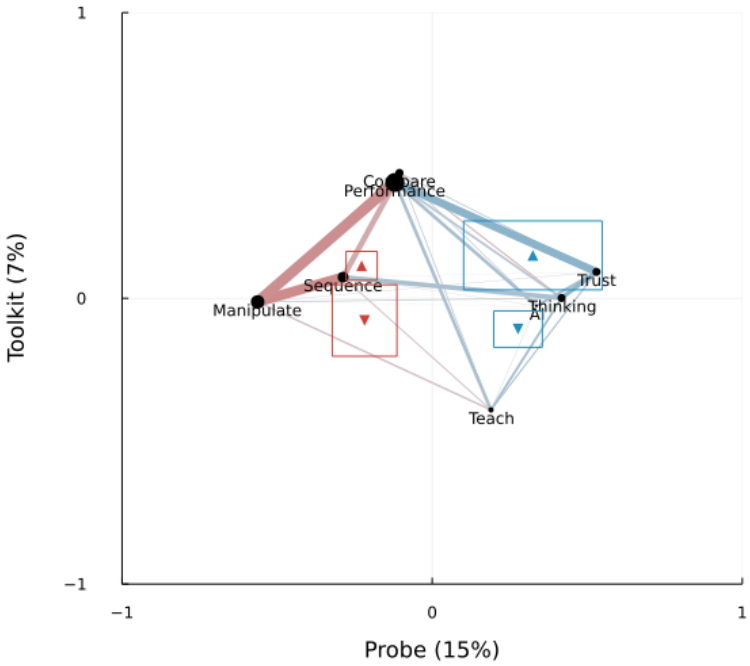

Fig. 5. Model 3 uses the x-axis to show the difference between discourse that occurred with a Probe, leading to more connections to *Trust* and *Thinking,* while the y axis depicts the impact of Toolkits leading to more connections to *Performance*.

6 Discussion

This study leverages the theory of teacher learning through participating in co-design processes to develop educational innovations. Specifically, we investigate how a co-design process can facilitate teacher thinking about assessment metrics, the role of artificial intelligence and algorithms, and data visualizations in game-based learning environments. To investigate this interplay between teacher's assessment literacy and co-design process, we took a quantitative ethnographic approach to analyzing the discourse data collected from the co-design process.

Related to RQ1, our results indicate that by participating in this co-design process, teachers gained a more sophisticated understanding of the role that artificial intelligence plays in game-based learning and assessment. That is, while they started with rather naive beliefs about how AI is being used in assessment and how they trust it, at the end of this process, they were expressing more critical views about how these algorithms are being created and used in educational technology. In addition, they demonstrated how they now think about what makes for trustworthy algorithms in relation to their own teaching practices and beliefs.

Related to RQ2, our results indicate different affordances of probes, toolkits, and prototypes to support the co-design process. Our finding is also closely aligned with the existing literature of co-designing with teachers who do not hold technical or design expertise. That is, probes allowed teachers to envision innovative forms of assessment in game environments beyond their current practices, and toolkits were helpful for teachers to create tangible artifacts that better reflect their desires and needs to support their students.

Our findings provide a few implications for the field of game-based learning and assessment. First, in contrast to the common practice of developing learning analytics models and algorithms without involving teachers, co-design methods can be used to provide a creative process that engage teachers to generate metrics and algorithms that they can make sense with and trust, ultimately increasing validity of the analytics. Second, the use of data in game-based learning environments is a powerful link that helps teacher to think about learning beyond scores and contents, and the field needs to thoughtfully approach analytics and data dashboards with the goal of teachers' capacity building.

In relation to QE, this work demonstrates how QE methods can be used to address the challenges related to unpacking the dynamic and iterative nature of co-design processes, not just in relation to the product development, but how a co-design process influences the participants as well. This is significant, especially given the emerging efforts to engage practitioners meaningfully in development of learning analytics [16] as teachers and students in more educational research [17].

7 References

1. Clark, D.B., Tanner-Smith, E.E., Killingsworth, S.S.: Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis. Review of Educational Research. 86, 79–122 (2016). https://doi.org/10.3102/0034654315582065.
2. Shute, V.J., Masduki, I., Donmez, O.: Conceptual Framework for Modeling, Assessing and Supporting Competencies within Game Environments. 1–25 (2011).
3. Lodge, J.M., Horvath, J.C., Corrin, L.: Learning analytics in the classroom: Translating learning analytics research for teachers. Routledge (2018).
4. Sanders, E.B.-N., Stappers, P.J.: Co-creation and the new landscapes of design. Co-design. 4, 5–18 (2008).
5. Sanders, E.B.-N., Stappers, P.J.: Probes, toolkits and prototypes: three approaches to making in codesigning. CoDesign. 10, 5–14 (2014).
6. Kumar, V.: 101 design methods: A structured approach for driving innovation in your organization. John Wiley & Sons (2012).
7. Gravemeijer, K., van Eerde, D.: Design research as a means for building a knowledge base for teachers and teaching in mathematics education. The elementary school journal. 109, 510–524 (2009).
8. Voogt, J., Laferrière, T., Breuleux, A., Itow, R., Hickey, D., McKenney, S.: Collaborative design as a form of professional development: in the context of curriculum reform. In: 2015 Annual Meeting of the American Educational Research

Association: Toward Justice: Culture, Language, and Heritage in Education Research and Praxis (2015).
9. Tsai, Y.-S., Gasevic, D.: Learning analytics in higher education---challenges and policies: a review of eight learning analytics policies. In: Proceedings of the seventh international learning analytics & knowledge conference. pp. 233–242 (2017).
10. Luckin, R.: Machine Learning and Human Intelligence: The future of education for the 21st century. ERIC (2018).
11. Rudin, C.: Algorithms for interpretable machine learning. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1519–1519 (2014).
12. Kim, Y.J., Lin, G., Ruipérez-Valiente, J.A.: Expanding Teacher Assessment Literacy with the Use of Data Visualizations in Game-Based Assessment. In: Visualizations and Dashboards for Learning Analytics. pp. 399–419. Springer (2021).
13. Marquart, C., Swiecki, Z., Eagan, B., Shaffer, D.W.: ncodeR: Techniques for automated classifiers [R package]. (2019).
14. Brohinsky, J., Marquart, C., Wang, J., Ruis, A.R., Shaffer, D.W.: Trajectories in Epistemic Network Analysis. In: International Conference on Quantitative Ethnography. ICQE (2021).
15. Knowles, M., Shaffer, D.W.: Hierarchical epistemic network analysis. In: Second International Conference on Quantitative Ethnography: Conference Proceedings Supplement. ICQE (2021).
16. Buckingham Shum, S., Crick, R.D.: Learning Analytics for 21st Century Competencies. Journal of Learning Analytics. 3, 6–21 (2016).
17. Ahn, J., Campos, F., Hays, M., DiGiacombo, D.: Designing in Context: Reaching Beyond Usability in Learning Analytics Dashboard Design. JLA. 6, 70–85 (2019). https://doi.org/10.18608/jla.2019.62.5.