Integrating Dynamic Supports into an Equity Teaching Simulation to Promote Equity Mindsets

G. R. Marvez, Tianyuan Zheng, Joshua Littenberg-Tobias, Garron Hillaire, Sara O'Brien, Justin Reich
Massachusetts Institute of Technology

ABSTRACT

Implementing high-quality professional learning on diversity, equity, and inclusion (DEI) issues is a massive scaling challenge. Integrating dynamic support using natural language processing (NLP) into equity teaching simulations may allow for more responsive, personalized training in this field. In this study, we trained machine learning models on participants' text responses in an equity teaching simulation (494 users; 988 responses) to detect certain text features related to equity. We then integrated these models into the simulation to provide dynamic supports to users during the simulation. In a pilot study (N = 13), we found users largely thought the feedback was accurate and incorporated the feedback in subsequent simulation responses. Future work will explore replicating these results with larger and more representative samples.

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**.

KEYWORDS

digital simulations, teacher education, natural language processing

ACM Reference Format:
G. R. Marvez, Tianyuan Zheng, Joshua Littenberg-Tobias, Garron Hillaire, Sara O'Brien, Justin Reich. 2022. Integrating Dynamic Supports into an Equity Teaching Simulation to Promote Equity Mindsets. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22), June 1–3, 2022, New York City, NY, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3491140.3528327

1 INTRODUCTION

Over the last few years, there has been a dramatic rise in interest in incorporating diversity, equity, and inclusion (DEI) training into teacher preparation and professional development. However, with over 3.5 million public school teachers in the country [11], implementing high-quality DEI professional learning poses a massive scaling challenge. Given the importance of DEI issues in educational settings, finding a method for delivering high-impact, low-cost, scalable teacher professional learning is an important policy goal.

Digital clinical simulations (DCS) offer a potentially promising means of implementing scalable online DEI training for educators. Digital clinical simulations employ rich digital media to immerse participants in a specific educational scenario and prompt participants to make improvisational instructional decisions at key moments in the scenario [9][13][14]. These simulations allow participants to consider teaching scenarios from an equity perspective within low-stakes scenarios where there are opportunities for reflection and revision [1][20][22]. Initial research into these simulations suggests that participants who engage in courses that embed these simulations develop more equitable mindsets about teaching [2][16].

One limitation of the first generation of these simulations is they did not provide dynamic support within the simulations themselves. Dynamic support using natural language processing (NLP) would allow participants to receive personalized feedback based on their individual responses [4]. This could be used, for example, to highlight specific aspects of the scenario that participants may have not noticed in their initial responses [10]. It could also be used to provoke cognitive dissonance in participants by noting actions that are inconsistent with their stated beliefs [15].

In this study, we present results from a pilot implementation integrating dynamic supports into an existing digital clinical simulation called Jeremy's Journal. We labeled responses from a corpus 988 human labeled responses from 494 users using binary classifiers to indicate the presence of key elements. We then trained machine learning models on this data to develop text classifiers that detected whether or not these elements were present in simulation text responses. We then integrated these classifiers into the existing Jeremy's Journal simulation using conditional statements to provide dynamic support based on participants' responses. We report the results from a pilot implementation of these supports with (N = 13) who participated in a workshop in February 2022. We address the following research questions:

(1) To what extent did participants perceive the classification of their responses as accurate and the feedback in the dynamic supports as useful?
(2) To what extent did participants change their behavior in the simulation immediately in response to the dynamic feedback?
(3) To what extent did participants' responses change on subsequent prompts in the simulation after receiving the dynamic feedback?

2 BACKGROUND AND CONTEXT

Approximations of practice are low-stakes opportunities to engage in simplified versions of real-life teaching situations [12]. Digital clinical simulations serve as approximations of practice by presenting a hypothetical teaching situation, such as a student misinterpreting a math problem, and asking them to improvise responses [14]. Simulations have been used as approximations of practice in a number of areas such such as improving teachers' questioning strategies [6][23], responses to student ideas [3], and teaching class discussion management [14]. Bywater et. al. (2019) devised a

L@S '22, June 01–03, 2022, New York, NY

ACM ISBN 978-1-4503-9158-0/22/06.
https://doi.org/10.1145/3491140.3528327

simulation tool called Teaching Responding Tool (TRT) to provide recommendations as instructor participants respond to students' explanations of mathematical concepts. The study coded the instructor participants' raw responses on a numeric scale from 0 to 2, representing the extent to which the instructor's feedback addressed a students' academic progress. As a direct result of TRT, most instructors reported shortened responses per explanation; furthermore, compared to non-TRT counterparts, their responses using TRT more accurately addressed students' learning status and suggested next steps.

Using digital clinical simulations, approximations of practice may be an effective method for helping teachers better understand how to break down barriers to learning for students from marginalized groups. Research on DEI efforts on teacher education has found that teachers benefit from opportunities to rehearse challenging situations and reflect on the reasoning behind their decisions [5][7][18]. Currently, many digital DEI simulations rely primarily on self-reflection [1]. Although self-reflection can be an effective tool, it is limited by the participant's own understanding of the situation. Some studies have found that simulation participants may retroactively justify their actions by framing it in equity language even if it does not match their own behavior [8][21]. For example, a teacher might justify harshly disciplining a student of color because they are upholding "high expectations." Integrating personalized feedback using NLP in simulations on DEI issues would provide participants with an alternative perspective on their behavior, which may cause them to question some of their original assumptions.

3 METHODS

3.1 Collecting Data for Labeling

The data that was used to train the NLP classifiers was collected from a simulation called Jeremy's Journal. Participants assume the role of a middle school English teacher who has a student named Jeremy. Jeremy is an outgoing student who faces a series of academic, personal, and health challenges during a week. Using images and text descriptions, participants are presented with various vignettes and asked how they would respond in the moment. At the end of the week, Jeremy asked to be excused from the required quiz. The data we used came from an implementation of Jeremy's Journal within a massive open online course (MOOC) on equity in education that was run from January - August 2021. Although the course enrolled 5,458 participants, we focused on participants who completed the Jeremy's Journal simulation and consented to participate in the research (N = 494). Of this sub-sample, 68% identified as female, 62% had a advanced degree, and 66% identified as native English speakers. The majority of participants reported working in K-12 schools (63%).

3.2 Labeling Data

Labeling was conducted by three raters with 20% of all texts (N = 197) randomly sampled to assess inter-rater reliability. Reliability was assessed at the beginning and middle of the rating process to ensure consistency in rating. Inter-rater reliability was generally good across all rater combinations with Cohen's kappa between 0.50-0.56, similar to what has been reported in previous research on similar types of tasks [17][3]. This labeled data then served as the "ground-truth" data for training our models.

We labeled 988 simulation responses from participants in the MOOC. For each set of participant responses, we devised at least six indicator variables representing potential criteria that responses may or may not satisfy. Examples of such variables include jeremy _effort, which is evaluated as "yes" if the response attributes Jeremy's subpar performance to lack of effort, lack of focus or being distracted, or negative attitude, and "no" otherwise. After close examination of each variable's description, a human rater reads each response and checks if it satisfies the variables' criteria by marking values 1 or 0 for "yes" or "no" respectively. This process is repeated by several human raters to improve rater inter-reliability.

We paid particular attention in restricting criteria to academic settings with variable definitions. Notably, when applicable, most variable definitions tend to accentuate academic observations and concerns, among other subtleties in a participant's instructional designs. For instance, in the dataset Enact_Monday_Wednesday, we defined change_for to evaluate as "yes" when a response proposes a change to the planned instruction based on student academic performance. If a response were to only mention instructional modifications such as "moving back a quiz" without observing the lacking academic performance of some subset of the class, the response would be rated as "0". Additionally, some variable definitions are specifically geared towards Jeremy's performance. In the case of learn_challenge, to satisfy the requirement, the response must identify potentially confusing components in the academic instructions and observe that Jeremy is particularly troubled. We impose those restrictions to improve the accuracy of our machine learning model predictions and the relevancy of our feedback to participants specific to their level of equity mindsets.

3.3 Natural Language Processing Models

To build the natural language processing models, we used the Scikit-learn Python package to create classifiers in Jupyter notebooks for our analysis [19]. Labeled participant responses were cleaned for punctuation, capitalization, and stopwords. These sanitized responses were then split into training and test data sets (80%/20%). We then examined six models for their performance: Random Forest, SVC (Linear), SVC (Sigmoid), SVC (Polynomial), SVC (RBF), and Decision Tree (Table 1). We selected the best model through the highest weighted average F1 score (Table 1). These models were pickled and added to a server that takes in participant data from Teacher Moments so that participants could receive personalized data.

In designing the personalized feedback that each participant receives, we designed six possible forms of feedback based on what the classifiers identified in the participant responses (Figure 1).

Participants saw one of these feedback questions with two simulation questions. For these questions, participants are asked how they might change their instruction based on Jeremy's work he turns in on Monday and Wednesday. Participants also learned that Jeremy has not been feeling well and has missed some class time before answering these questions.

3.4 Piloting Feedback

We piloted the personalized feedback in Jeremy's Journal with participants in a workshop organized by the authors. The goal of the workshop was connect teacher educators and learning analytics researchers, to together to author simulations, train classifiers using the data they collected, and use those classifiers to create dynamic supports in their simulations. Participants completed the simulation after they had learned about authoring the simulation and NLP, but before they learned about dynamic supports. Out the 12 participants who shared their current work position, 6 were learning analytics researchers, 3 were teacher educators, 3 were current K-12 teachers, and 1 identified as "other". Only participants who consented to participate in the research (N = 13) were included in the data analysis.

Table 1: F1 statistics for all models

Model	1	2	3	4	5	6
Random Forest	0.91	0.69	0.67	0.71	0.84	0.84
SVC (Linear)	0.94	0.60	0.67	0.70	0.83	0.80
SVC (Polynomial)	0.85	0.61	0.58	0.64	0.84	0.81
SVC (Sigmoid)	0.94	0.70	0.69	0.75	0.83	0.86
SVC (RBF)	0.91	0.66	0.69	0.68	0.84	0.82
Decision Tree	0.91	0.69	0.69	0.70	0.80	0.81

Note: Column labels are 1-feel_jeremy, 2-learn_challenge, 3-change_for, 4-more_some, 5-jeremy_mental, 6-jeremy_effort.

4 RESULTS

RQ.1: To what extent did participants perceive the classification of their responses as accurate and the feedback in the dynamic supports as useful?

Participants received feedback twice in the simulation. After receiving feedback, we asked participants, "Did you think your response was evaluated correctly?" Participants could choose one of three responses: It was totally wrong; it got some things right, but missed some things; or it was totally correct. Thirteen participants completed the first question and 11 completed the second. The models we deployed provided at least partially helpful feedback for 92% and 91% of participants in the first and second round of feedback (Table 2).

Feedback Round	Rating	N
First	It was totally correct	4
	It got some things right	8
	It was totally wrong	1
Second	It was totally correct	5
	It got some things right	5
	It was totally wrong	1

RQ.2: To what extent did participants change their behavior in the simulation immediately in response to the dynamic feedback?

In response to the feedback, participants reflected on their treatment of Jeremy and what barriers he may have been facing to complete his work. There were 16 total responses from participants to the personalized feedback, but two were left blank, leaving

Classifiers identified the response…	Feedback Received
Mentioned Jeremy's behavior	Based on your response, you have mentioned concerns about Jeremy's behavior in class. Why do you think he might act this way? Is there anything about your teaching that might be challenging for him?
Did not mention Jeremy's behavior, his learning challenges, or a need for a change in instruction	Based on your response, you may have not specifically mentioned Jeremy's academic needs and how you might support him. Did you notice anything in Jeremy's work that might make you think he didn't understand some part of the task? What kind of support might he need?
Mentioned his learning challenges and a need for a change in instruction	Based on your response, you may have mentioned Jeremy's academic needs, but not his social/emotional needs as a learner? What do you think might be challenging for him right now? What kinds of social/emotional support could you provide?
Mentioned Jeremy's social-emotional needs, but not his academic challenges	Based on your response, you may have mentioned Jeremy's something about how Jeremy's social/emotional needs but not his academic needs? . Did you notice anything in Jeremy's work that might make you think he didn't understand some part of the task? What kind of supports might he need?
Mentioned Jeremy's social-emotional needs and his academic challenges	Based on your response, you may have mentioned Jeremy's academic needs and his social/emotional needs as learning? What do you think might be challenging for him right now? How could you continue to support Jeremy?
Mentioned Jeremy's academic challenges, but not his social-emotional needs	Based on your response, you may have mentioned Jeremy's academic needs, but not his social/emotional needs as a learner. What do you think might be challenging for him right now? What kinds of social/emotional supports could you provide?"

Figure 1: The six forms of feedback participants could receive based on their response classification.

14 text responses. We analyzed if they successfully responded to the feedback by incorporating the feedback into an adjustment in their instruction, or in the way they interacted with Jeremy. For instance, if the personalized feedback suggested that the participant think about Jeremy's social-emotional needs, did the participant's reflection show they considered this aspect?

Nine of the 14 responses showed the participant reflected on an aspect of Jeremy's experience in the classroom that they did not consider (64%), four responses were confused about what problems Jeremy may be facing, and one response outright disagreed with the feedback received.

RQ.3: To what extent did participants' responses change on subsequent prompts in the simulation after receiving the dynamic feedback?

In the final aspect of the scenario, participants respond to Jeremy's concerns about taking the quiz on the same day as the rest of the class and explain their decision on whether he should take the quiz or not. Here, an equitable response would highlight that Jeremy should not take the quiz without modifications due to the academic and health challenges he has faced this week. Nine participants completed this part of the scenario.

In their Monday responses, five participants received feedback that they did not specifically mention an academic challenge Jeremy was facing, and four received feedback that they mentioned one

of Jeremy's academic struggles, but did not consider his social-emotional needs. In their Wednesday responses, two received feedback that they seemed concerned about Jeremy's behavior and what might be affecting his behavior, five received feedback that they did not specifically mention an academic challenge Jeremy was facing, and two received feedback that they mentioned one of Jeremy's academic struggles, but did not consider his social-emotional needs.

Two participants stated that he should take the quiz and seven said he should not take the quiz. The two participants who stated that they wanted Jeremy to take the quiz mentioned that they wanted to use the quiz as a benchmark to figure out what Jeremy understands. These two participants had received feedback to consider the reasons for Jeremy's behavior, his academic concerns, and his social-emotional needs. Of the seven who stated he should not take the quiz, five stated concerns about his mental health, and three mentioned an academic concern about the quiz. These seven participants also saw a mix of feedback on Jeremy's behavior, academic struggles, and social-emotional needs.

5 CONCLUSIONS AND FUTURE WORK

In this pilot run of responsive in-the-moment feedback for participants in a digital clinical simulation, we find that using traditional machine learning models trained on simulation data can help provide participants guidance on their performance during simulations. Additionally, participants found this feedback to be largely accurate and integrated the feedback suggestions into their later performance in the scenario. Future work will be needed to adjust the parameters of our models to provide a more accurate classification and to redesign the feedback participants receive to ensure that they have adequate reflection opportunities. And while early evidence suggests that the supports embedded in the simulation are influential, future work should consider A/B testing support variations to identify optimal supports in terms of changing immediate behavior as well as the long-term effects on behavior change. To conduct such an A/B test, we would like to scale up this study with a larger group of participants. In this demo for Learning at Scale, participants will be able to interact with the natural language processing feedback system in the scenario, and provide feedback to the designers on ways to improve the user experience. This material is based on work supported by the National Science Foundation under grant number 1917668 and the Bill and Melinda Gates Foundation.

REFERENCES

[1] Elizabeth Borneman, Joshua Littenberg-Tobias, and Justin Reich. 2020. Developing digital clinical simulations for large-scale settings on diversity, equity, and inclusion: Design considerations for effective implementation at scale. In *Proceedings of the Seventh ACM Conference on Learning @ Scale (L@S '20)*. Association for Computing Machinery, New York, NY, USA, 373–376. https://doi.org/10.1145/3386527.3405947

[2] Christopher J. Buttimer, Joshua Littenberg-Tobias, and Justin Reich. 2022. Designing online professional learning to support educators to teach for equity during covid and black lives matter. *AERA Open* 8 (Jan. 2022), 23328584211067789. https://doi.org/10.1177/23328584211067789 Publisher: SAGE Publications Inc.

[3] James P. Bywater, Jennifer L. Chiu, James Hong, and Vidhya Sankaranarayanan. 2019. The Teacher Responding Tool: Scaffolding the teacher practice of responding to student ideas in mathematics classrooms. *Computers & Education* 139 (Oct. 2019), 16–30. https://doi.org/10.1016/j.compedu.2019.05.004

[4] Xieling Chen, Haoran Xie, Di Zou, and Gwo-Jen Hwang. 2020. Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence* 1 (Jan. 2020), 100002. https://doi.org/10.1016/j.caeai.2020.100002

[5] Sauro Civitillo, Linda P. Juang, and Maja K. Schachner. 2018. Challenging beliefs about cultural diversity in education: A synthesis and critical review of trainings with pre-service teachers. *Educational Research Review* 24 (June 2018), 67–83. https://doi.org/10.1016/j.edurev.2018.01.003 Publisher: Elsevier Ltd.

[6] Debajyoti Datta, Maria Phillips, James P. Bywater, Jennifer Chiu, Ginger S. Watson, Laura E. Barnes, and Donald E. Brown. 2021. Evaluation of mathematical questioning strategies using data collected through weak supervision. *arXiv:2112.00985 [cs]* (Dec. 2021). http://arxiv.org/abs/2112.00985 arXiv: 2112.00985.

[7] Michael Domínguez. 2020. Cultivating epistemic disobedience: exploring the possibilities of a decolonial practice-based teacher education. *Journal of Teacher Education* (2020). https://doi.org/10.1177/0022487120978152

[8] Benjamin Dotger and Christine Ashby. 2010. Exposing conditional inclusive ideologies through simulated interactions. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children* 33, 2 (May 2010), 114–130. https://doi.org/10.1177/0888406409357541

[9] Ritam Dutt, Garron Hillaire, Alison Fang, Laura Larke, Carolyn Rosé, and Justin Reich. 2021. Investigating adoption and collaboration with digital clinical simulations by teacher educators. Association for the Advancement of Computing in Education (AACE), 1209–1217. https://www.learntechlib.org/primary/p/219277/

[10] Aria Eppinger, G.R. Marvez, Josh Littenberg-Tobias, Garron Hillaire, Sydney Dell, and Jusitn Reich. 2022. From avoidant to aware: Automating feedback in simulations on equity in computer science education. San Diego, CA, US.

[11] National Center for Education Statistics. 2020. Characteristics of public school teachers. https://nces.ed.gov/programs/coe/indicator_clr.asp

[12] Pamela Grossman, Christa Compton, Danielle Igra, Matthew Ronfeldt, Emily Shahan, and Peter Williamson. 2009. Teaching practice: A cross-professional perspective. *Teachers College Record* 111, 9 (Sept. 2009), 2055–2100. https://www.tcrecord.org/Content.asp?ContentId=15018

[13] Garron Hillaire, Laura Larke, and Justin Reich. 2020. Digital Storytelling through Authoring Simulations with Teacher Moments. Association for the Advancement of Computing in Education (AACE), 1756–1765. https://www.learntechlib.org/primary/p/215950/

[14] Sarah J. Kaka, Joshua Littenberg-Tobias, Taylor Kessner, Anthony Tuf Francis, Katrina Kennett, G. Marvez, and Justin Reich. 2021. Digital simulations as approximations of practice: preparing preservice teachers to facilitate whole-class discussions of controversial issues. *Journal of Technology and Teacher Education* 29, 1 (2021), 67–90. https://www.learntechlib.org/primary/p/218711/ Publisher: Society for Information Technology & Teacher Education.

[15] Laura R Larke, Garron Hillaire, Hao Chen, Ritam Dutt, Carolyn Penstein Rose, and Justin Reich. 2020. Cognitive Dissonance and Equity: Designing Digital Simulations for K-12 Computer Science Teacher Education. 2.

[16] Joshua Littenberg-Tobias, Elizabeth Borneman, and Justin Reich. 2021. Measuring equity-promoting behaviors in digital teaching simulations: A topic modeling approach. *AERA Open* 7 (Jan. 2021), 23328584211045685. https://doi.org/10.1177/23328584211045685 Publisher: SAGE Publications Inc.

[17] Ou Lydia Liu, Chris Brew, John Blackmore, Libby Gerard, Jacquie Madhok, and Marcia C. Linn. 2014. Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles. *Educational Measurement: Issues and Practice* 33, 2 (2014), 19–28. https://doi.org/10.1111/emip.12028 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12028.

[18] Hillary Parkhouse, Chu Yi Lu, and Virginia R. Massaro. 2019. Multicultural education professional development: A review of the literature. *Review of Educational Research* 89, 3 (June 2019), 416–458. https://doi.org/10.3102/0034654319840359 Publisher: American Educational Research Association.

[19] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* (2011), 6.

[20] Kevin Robinson, Keyarash Jahanian, and Justin Reich. 2018. Using online practice spaces to investigate challenges in enacting principles of equitable computer science teaching. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE '18)*. Association for Computing Machinery, New York, NY, USA, 882–887. https://doi.org/10.1145/3159450.3159503

[21] Elizabeth A. Self and Barbara S. Stengel. 2020. *Toward anti-oppressive teaching: Designing and using simulated encounters.* Harvard Education Press, Cambridge, Massachusetts.

[22] Meredith Thompson, Kesiena Owho-Ovuakporie, Kevin Robinson, Yoon Jeon Kim, Rachel Slama, and Justin Reich. 2019. Teacher moments: a digital simulation for preservice teachers to approximate parent–teacher conversations. *Journal of Digital Learning in Teacher Education* 35, 3 (July 2019), 144–164. https://doi.org/10.1080/21532974.2019.1587727

[23] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a Few Examples: A Survey on Few-shot Learning. *Comput. Surveys* 53, 3 (June 2020), 63:1–63:34. https://doi.org/10.1145/3386252