Piloting a Vignettes Assessment to Measure K-5 CS Teacher Proficiencies and Growth

Joseph C. Tise
joe@csedresearch.org
Institute for Advancing Computing Education
Winchester, VA, USA

Monica M. McGill
monica@csedresearch.org
Institute for Advancing Computing Education
Peoria, IL, USA

Vicky Sedgwick
visionsbyvicky@gmail.com
Computer Science Teachers Association
Canoga Park, CA, USA

Laycee Thigpen
lthigpen44@gmail.com
Institute for Advancing Computing Education
Lakeland, FL, USA

Amanda Bell
amanda.bell@csteachers.org
Computer Science Teachers Association
Methuen, MA, USA

Abstract

Background. The Standards for Computer Science (CS) Teachers include indicators related to classroom practices. To assess teacher proficiency related to these indicators at scale, we created and pilot-tested a vignette-based measure of K-5 CS teacher proficiencies related to Standards 2, 4, and 5.

Research Questions. Our two research questions were: 1) *How difficult did teachers find the new measure, and how effectively did each item discriminate between high- and low-performers?* 2) *Which teacher characteristics predict scores on the new measure?*

Methodology. We developed three vignettes and aligned associated questions to Standards 2, 4, and 5. We conducted cognitive interviews with teachers, then piloted the instrument with 111 U.S. K-5 teachers. Using classical test theory, we assessed its reliability and validity as well as each item's difficulty and discrimination values.

Key Findings. Scores on the measure were approximately normally distributed. Item difficulties ranged from .46 (somewhat difficult) to .95 (very easy). Item discrimination values ranged from .16 to .48. Cronbach's alpha ($\alpha = .66$) indicated the measure could be improved to increase reliability. Scores on the measure were positively correlated with teachers' reported teaching awards, but were not predicted by any of the independent variables.

Implications. This new measure of teacher proficiencies shows mixed psychometric qualities, and additional revisions to the items are warranted. Once finalized, this measure could potentially be used by practitioners to identify strengths and growth areas for future professional development.

CCS Concepts

• **Social and professional topics → K-12 education**; **Computer science education**; **Computing education programs**; **Professional topics**; **Computing education**;

SIGCSE TS 2026, St. Louis, MO, USA

ACM ISBN 979-8-4007-2256-1/2026/02
https://doi.org/10.1145/3770762.3772604

Keywords

Assessment, CS Teacher PD, Equity, Vignettes, CSTA Standards for Teachers, Elementary CS Teachers

ACM Reference Format:
Joseph C. Tise, Monica M. McGill, Vicky Sedgwick, Laycee Thigpen, and Amanda Bell. 2026. Piloting a Vignettes Assessment to Measure K-5 CS Teacher Proficiencies and Growth. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.1 (SIGCSE TS 2026), February 18–21, 2026, St. Louis, MO, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3770762.3772604

1 Introduction

The need for educators to continually grow in their content expertise, especially in emerging fields like computer science (CS), underscores the importance of reliable assessment tools to follow professional development (PD) and support teacher growth. To further this process, the Computer Science Teachers Association (CSTA) *Standards for CS Teachers* defines five Standards [4] that correspond with *CS Knowledge and Content*, *Equity and Inclusion*, *Professional Growth and Identity*, *Instructional Design*, and *Classroom Practice*. Released in 2019, these standards support a philosophy that every student can learn CS through teachers who practice inclusive instruction and continuous professional growth. It is critical for CS teachers to engage in inclusive instruction to support each and every student in their classroom. The standards have been used to guide curriculum for teacher PD [16] and also serve as a basis for self-reflection from teachers [3].

We developed and piloted a set of vignettes for K-5 teachers to assess teacher growth across three of the standards related to classroom practice (2: Equity and Inclusion; 4: Instructional Design; and 5: Classroom Practice). This follows a series of scales we developed and pilot tested in 2022 and 2023 [11, 18].

To assess the psychometric quality and provide preliminary evidence of validity for our new measure, we posed the following research questions:

(1) *How difficult did teachers find the new measure, and how effectively did each item discriminate between high- and low-performers?*
(2) *Which teacher characteristics predict scores on the new measure?*

2 Background

Similar to growth in students, measuring growth in teachers can alleviate stressors that are associated with summative feedback [9]. As noted by Alicea et al., practical measures should be useful (actionable data for practitioners); minimally burdensome to collect and analyze, but predictive of longer-term outcomes; and consequential in a way that it is connected to experiences that matter for learning and to outcomes we care about for students–especially those who have been traditionally underserved [1].

Often, researcher-developed instruments for teacher growth miss one or more of these characteristics and they are also often not aligned to PD providers' needs [1, 21]. This leaves an opportunity to create such measures that can assess teachers' growth and help them identify specific areas of development on which to focus.

2.1 Measuring Teacher Growth

In 2022, we streamlined the CSTA standards by condensing 29 indicators across Standards 2-5 into 18 rubric items to create a diagnostic tool [11]. This tool allowed teachers to submit artifacts like lesson plans or videos and explain how these materials aligned with each rubric item. We piloted the measurement with 24 teachers in the U.S. who used the diagnostic tool before participating in a CSTA Teacher CS PD week. After incorporating their feedback, an additional 29 teachers used the assessment. This pilot phase revealed that the assessment reports were helpful to teachers, but they had several issues, such as cognitive overload given the instrument's length (roughly three hours to collect all of the artifacts and enter them into the instrument with justifications) and complexity.

As we refined the diagnostic tool, our goal was to make it more user-friendly while maintaining its effectiveness. We reduced the 18 rubric items to 14 and began exploring whether directly asking teachers about their PD needs would yield different or more useful results compared to the diagnostic ratings. While they ultimately did not [18], the amount of time to take the assessment was cut in half. We piloted this self-report measure with 66 teachers, and feedback from teachers was overwhelmingly positive. However, given the amount of data to score, our team trained a team of reviewers and facilitated their reviews, an effort that took over 400 hours. It thus became abundantly evident that although scaling such a measure is desirable, it is not practical or minimally burdensome.

2.2 Measuring Growth via Vignettes

Developing a multiple-choice, vignettes-based assessment could be a practical alternative. Vignettes have been used through teacher training and have become a common and acceptable method for measuring pedagogical content knowledge [15, 17**?**]. Vignettes offer a way to assess teacher proficiency by presenting authentic classroom scenarios through a short story [**?**]. Vignettes can offer insight into what teachers know about authentic classroom settings beyond their own, offering details about the setting and context for the items [**?**]. Unlike traditional forms of assessment, vignettes bridge a gap between classroom observation (which is challenging to collect at scale) and instruments that rely on self-reported accounts by teachers. Vignettes require teachers to draw upon their knowledge, experiences, and judgment to make decisions relevant to the vignettes. Further, though vignettes cannot replace classroom observations of teachers, a method that has its own limitations [6, 13], they offer a way for teachers to reflect on a described classroom environment and then reflect and answer questions about the hypothetical teacher's behavior.

Within CS, vignettes have been explored to measure CS pedagogical content knowledge [15, 20]. Given this precedent, we pivoted to an assessment measure that was based on vignettes. We chose K-5 teachers for this initial pilot for assessing teacher knowledge against three standards (2, 4, and 5), since Standard 1 is focused on CS content knowledge which requires a different type of assessment and Standard 3 is focused on teacher professional growth (such as creating a CS professional development plan). The indicators for each standard are shown in Table 1.

2.3 Defining Teacher Proficiency

Creating a measure to assess teacher proficiency against any standard requires the ability to discern between the responses from expert teachers and novice teachers. While there is a dearth of research on differentiating teacher expertise across the CSTA Standards for CS Teachers, research to discern expert from novice teachers in related fields like mathematics and science has been done with interesting results. Kosel et al. compared the ability for in-service teachers ($n = 19$) and pre-service teachers ($n = 24$) to discern their judgment of surface cues (e.g., raising hands) and deep cues (e.g., student interest in the subject being taught) in a video that they watched [7]. Overall, in-service teachers experienced more judgment accuracy in their reasoning. In a different study by Kosel et al., the researchers measured differences in novice teachers and experienced teachers eye gaze during a regular class period using eye-tracking glasses that the participants wore during the lesson. Notably, they found experienced teachers' gaze index was lower than novice teachers' and they interpreted this data as "...more fine-graded organization of domain-specific knowledge allows them to fixate more rapidly and frequently in the classroom" [8, p. 267].

Researchers have operationalized the expert-novice continuum in numerous ways. Criteria that researchers have used often differ, but traditionally include one or more of the following:

- In-service vs pre-service [7]
- Average years of experience teaching (novice less 2 years; experienced more than 7 to 10 years) [8, 10, 19]
- Identified by school leaders or educational authorities [10, 19]
- Held licensure at a certain level [10]
- Recommendations from fellow teachers as competent or above-average teacher [19]
- Evidence of meaningful reflective practices focused on improving student learning (as measured by teachers' detailed approaches to and what they learned from reflection) [2]
- Evidence of effectiveness at facilitating student learning (as measured by pre- and post-tests of students) [2]

While some of these differentiating criterion relied on years of experience, experience is not equivalent to expertise–particularly when it comes to the CSTA teacher standards. Validation therefore requires more than experience; triangulation with other measures such as student outcomes, pedagogical confidence, self-reflection, and observed behaviors are needed.

Table 1: Standards 2, 4, and 5 with their indicators [4].

Standard	Title	Indicators
2	**Equity and Inclusion** Teachers proactively advocate for equity and inclusion in the CS classroom.	**2a.** Examine issues of equity in CS—structural, social, and psychological barriers—and reflect on personal teaching contexts. **2b.** Minimize threats to inclusion by addressing unconscious bias and stereotype threat. **2c.** Represent diverse perspectives through inclusive curriculum and instruction. **2d.** Use data to make decisions that improve equity in access, engagement, and participation. **2e.** Select and adapt accessible instructional materials for all learners.
4	**Instructional Design** Teachers design learning experiences using pedagogical content knowledge, tailored to diverse student needs.	**4a.** Analyze CS curricula for alignment, accuracy, completeness, and cultural relevance. **4b.** Develop standards-aligned learning experiences. **4c.** Design inclusive learning experiences using UDL and culturally relevant pedagogy. **4d.** Integrate real-world and cross-curricular connections into CS instruction. **4e.** Plan open-ended projects that have personal meaning to students. **4f.** Use CS-specific strategies to build understanding and address misconceptions. **4g.** Use assessments to inform instruction and support student learning.
5	**Classroom Practice** Teachers implement effective, responsive instruction that empowers learners.	**5a.** Use inquiry-based methods (e.g., PRIMM, peer instruction, POGIL). **5b.** Cultivate a safe, inclusive, and respectful classroom environment. **5c.** Facilitate student-centered learning experiences. **5d.** Use formative and summative assessments to support learning. **5e.** Manage classroom routines and technologies to enhance instruction.

3 Methodology

To answer our research questions, we conducted item analyses (via classical test theory) and multiple linear regression to determine which teacher characteristics (if any) predicted total scores on the vignettes measure.

3.1 Vignette Development

The vignettes were developed by a team of four people: three with education research backgrounds, one with extensive experience teaching K-5 CS, and one with early childhood education experience. One participated in drafting the Standards for Teachers. Our process included a continual review with two team members from the Computer Science Teachers Association (CSTA), who provided constructive feedback on the scenarios and responses.

Once we established what standards would be embedded into which vignette, we created an outline for the scenarios to capture unique classroom experiences (i.e., geographic locale, types of students, experience levels of the teachers). After discussing and refining as a group, the four developers developed the scenario and the questions for each standard in the first vignette. After completing the draft, we shared with the two CSTA members, who provided critical feedback. This process continued two more times until there was no more feedback. This process was then repeated for the remaining two vignettes.

Early in the process, the vignette developers created three areas on which each vignette would focus: planning a lesson, teaching the lesson, and assessing student knowledge of the lesson content. These three areas were created in the second year of this project and are described more in [18]. The vignette developers then created a framework to guide answer choice development. Each question included one correct answer and three distractors as follows:

- Correct answer (ideal action): Items which demonstrate clearly good practice (generally and in CS ed); aligned with the target standard
- Distractor (good): Items that may be good practice generally, but which do not fully align with the assessed standard
- Distractor (neutral): Items that won't hurt student learning, but which also likely won't benefit them either
- Distractor (detrimental): Items that would potentially harm student learning

We have opted to not share the vignettes publicly because they are still under development and are an assessment. Readers may contact one of the authors on this study for a copy of the vignettes.

3.1.1 Cognitive Interviews. After the vignettes were completed, one of the researchers conducted cognitive interviews with four experienced CS teachers and one new CS teacher. All teachers (experienced and new to CS) had several years of prior teaching experience, and one expert CS teacher had a background in CS curriculum development. The researcher met with each teacher for a 1-hour cognitive interview for each of the three vignettes. During each session, the researcher walked the teacher through each main component of the vignette (background scenario and each follow-up question), and stopped after each component to ask questions regarding clarity and bias. For each follow-up question, the teacher indicated their answer choice and then provided their reasoning for selecting their answer choice and their reasoning for not selecting the alternatives. After all cognitive interviews were complete, the research team synthesized this feedback and cross-referenced it with each teacher's performance on the items to determine if the teachers' reasoning for each answer choice was aligned with our intentions for each answer choice. This process led us to significantly revise four items across the three vignettes.

3.2 Discerning Novice from Expert

Teachers can fall into one of four categories: new teacher with CS experience, new teacher and new to CS, experienced teacher but new to CS, and experienced CS teacher [3], making it difficult to distinguish between novice and expert CS teachers. To properly assess the instrument, we needed to define what constitutes a CS teacher who 1) practices the standards proficiently, and 2) does not practice the standards proficiently. Given the dearth of research in this area, our team discussed how we could differentiate between the two. To this end, we collected a wide range of data, including:

- Years of teaching experience
- Years of CS teaching experience
- On average, how often they teach CS (e.g., daily, weekly, monthly)
- Familiarity with Standards 2, 4, and 5 (self-rated, 1-7)
- Number of teaching awards received (self-reported, numeric)
- Number of CS teaching awards received (self-reported, numeric)
- Participation in CSTA activities (e.g., conference, regional chapter, webinars)
- Number of CS PD participated in (self-reported, numeric)
- Self-reported use of classroom practices aligned with the Standards
- Self-efficacy for teaching CS (6 items, rated 1-7)

3.3 Participants

Teachers ($N = 111$) were recruited for CSPDWeek through multiple channels and marketing campaigns. Recruitment began in October and continued into the following spring. The planning team reached out to attendees from prior CSTA PD events to encourage them to apply. The team also sent out information about the event to our national and local membership email lists, school administrators, school counselors, and state Department of Education partners. Additionally, we provided marketing materials for the organizations facilitating workshops at CSTA PD events to recruit within their networks. Incentives included a $1,080 stipend for completing all pre-work and attending the full five days of PD, along with a completion certificate and continuing education hours. Participants were also eligible for a second stipend of up to $480 for attending follow-up PD sessions during the subsequent school year. As part of the pre-work, teachers were asked to complete the Vignettes assessment.

Descriptive statistics pertaining to their teaching and PD activities are in Table 2. Teachers in this sample were experienced on average ($M_{years\ teaching} = 15.50, SD = 9.28, Median = 15, Min = 0. Max = 35$), but were relatively new to teaching CS, on average ($M_{years\ teaching\ CS} = 3.90, SD = 3.89, Median = 3, Min = 0. Max = 25$). Further, these teachers participated in 5 to 6 CS PD opportunities on average across their careers.

3.4 Analyses

Fifteen responses were removed due to completely missing data, yielding a final analytic sample of 111. For each item, we report discrimination and difficulty values. An item's discrimination value indicates how well the item can differentiate between higher- and lower-performers. Statistically, it is the point-biserial correlation between participants' scores on the item (i.e., correct or incorrect) and their sum scores of all other test items. Thus, it can range in value from -1 to +1; higher positive values are desirable [5, 14]. An item with a discrimination value of 1 indicates every person who answered it correctly ended up scoring 100% on the other items, while everyone who answered it incorrectly ended up scoring 0% on the other items. Finally, an item's difficulty value indicates the proportion of the sample that correctly answered the item. It is thus expressed simply as a percentage.

For RQ2, we analyzed a multiple linear regression model, with six independent variables (teaching experience, CS teaching experience, receipt of teaching awards [Yes/No], receipt of CS teaching awards [Yes/No], participation in a professional learning community in the last year [PLC; Yes/No], participation in other CS PD opportunities in the last year [Yes/No], number of CS activities participated in, number of PD opportunities participated in ever, and self-assessed proficiencies aligned with the CSTA teacher standards [scale score]) predicting our primary dependent variable (total score on the vignettes measure). The Bonferroni correction was applied to account for family-wise Type I error.

Table 2: Demographic data of participants

Variable	*N*	%
Teach CS at least once per week	49	44.14
Participated in a CS professional learning community in the last year	40	36.04
Participated in other CS PD (workshops, conferences)	63	56.76
Asian or Asian American	1	0.90
Black or African American	2	1.80
Hispanic or Latinx	1	0.90
White	101	90.99
Multiple race/ethnicities	2	1.80
Race/ethnicity not reported	4	3.60
Women	97	87.39
Men	11	9.91
Gender not reported	3	2.70
Has a disability	3	2.70

4 Results

4.1 RQ1: Difficulty and Discrimination

Participants scored relatively highly, on average, on the measure ($M = 76.36\%, SD = 12.47\%, Median = 78.57\%, Min = 35.71\%, Max = 97.62\%$). These descriptive statistics are similar to those obtained from a measure of teachers knowledge pertaining specifically to CSTA Teacher Standard 1 [12]. The scale demonstrated less-than-ideal reliability ($\alpha = .55$). This may have been due to numerous factors, including the dichotomously-scored nature of the items [**?**]. Still, additional refinement of the measure is warranted. The average item difficulty was 0.76, indicating that, on average, 76% of participants correctly answered a given item.

Table 3: Average item difficulty across the items and indicators within each Standard.

Vignette # Focus Area	Item	Indicator Assessed	Difficulty	Discrimination
1 Plan	1	4a, 4b	0.79	0.31
1 Plan	2	4a	0.81	0.36
1 Plan	3	4a, 4d	0.86	0.35
1 Plan	4	4f	0.50	0.16
1 Teach	5	5b	0.83	0.36
1 Teach	6	2a	0.93	0.32
1 Asesss	7	2d	0.94	0.29
1 Assess	8	4g	0.75	0.37
1 Assess	9	4f	0.68	0.41
2 Plan	10	4e	0.82	0.46
2 Plan	11	2c	0.68	0.17
2 Plan	12	4c	0.62	0.43
2 Plan	13	2b	0.78	0.33
2 Teach	14	5d	0.75	0.40
2 Teach	15	5e	0.88	0.17
2 Assess	16	5f	0.95	0.46
3 Plan	17	4a	0.82	0.27
3 Plan	18	5c	0.86	0.27
3 Plan	19	2e	0.75	0.48
3 Teach	20	5a	0.46	0.18
3 Assess	21	2d	0.58	0.31

4.2 RQ2: Predicting Total Scores on the Measure

Correlations among all variables included in research question 2 are presented in Table 4. As seen, only two variables correlated significantly with teachers' total scores on the vignettes measure. Specifically, teachers who had won at least one teaching award tended to have higher scores on the measure, while teachers who had participated in at least one professional learning community in the past year tended to have lower scores.

Results from the multiple linear regression model indicated that none of the independent variables predicted teachers' total scores (Table 5). Figuring that more-expert teachers would have higher self-efficacy to teach CS than novice teachers, we re-ran the same model adding self-efficacy as a predictor of total score. Due to a technical error in the survey software, only half ($n = 48$) participants reported self-efficacy data. Self-efficacy did not statistically significantly predict scores on the measure ($\beta = -1.33$, $t = -0.76$, $p = .999$), and results pertaining to the other predictors were not changed.

5 Discussion

This study presents preliminary psychometric information and validity evidence for a new vignettes-based measure of K-5 CS teacher proficiencies related to Standards 2, 4, and 5. Measuring teacher proficiencies is difficult, and even more difficult at scale via multiple-choice tests. Still, we need valid and scalable assessments to accomplish myriad goals, such as providing formative feedback to teachers and generating reliable measures for use in research. Such was the overarching goal of this study.

Results from RQ1 indicated that most items were relatively easy for the present sample. Only two of the 21 items (9.5%) were answered correctly by half the sample or less, while 15 of the 21 (71%) were answered correctly by at least 75% of the sample. At the same time, most items were also moderately effective at discriminating between higher- and lower-performers. Still, the measure could be refined further and then re-administered to a larger and perhaps more diverse sample of teachers.

For RQ2, we tested a number of factors as we struggled to define criteria to differentiate between teachers who do and do not employ the practices in the Standards. While (*teaching awards* and *participation in a professional learning community* correlated with total scores on the measure, none of the independent variables predicted total scores. This differs from previous studies which have shown that years of teaching experience [8, 10, 19] can differentiate novice and expert teachers. But given the nature of our focus on Standards 2, 4, and 5, our experts needed to be uniquely attuned to classroom practices that supported those standards. Experienced teachers may or may not value inclusive teaching practices, for example, or practices that incorporate specific types of instructional design for teaching CS (i.e., using a variety of assessment measures).

We were surprised to see that the number of teaching awards correlated with scores ($r = .19$) while number of *CS* teaching awards did not. Upon further investigation into the data, we hypothesize these results could have manifested because only eight teachers (7% of our sample) reported any CS teaching awards, and all eight of those teachers reported only one award. Conversely, 27 teachers (24% of our sample) reported receiving at least one (general) teaching award, with some reporting two or even three awards. Thus, there was extremely limited variance present in the CS teaching awards variable, which likely influenced the correlation coefficient.

While we had hoped that teacher's self-reported level of their own classroom practices that are aligned with standards would correlate with scores on our measure, the fact that they did not mirrors results we have found in one of our previous studies [11, 18].

The somewhat contradictory finding of the two dichotomous variables may have emerged because the two significant correlations were border-line non-significant. Further, the dearth of significant predictors and correlates with scores on the measure could be explained in at least three ways. First, the variables that *should* be included in the nomological net (from a theoretical standpoint) with teacher proficiencies may not include the variables we assessed in this study. Given the nascent theoretical development on this front, we may have simply not identified the proper correlates/indicators of true teacher proficiency aligned to the CSTA Teacher Standards.

Second, if the variables we investigated were to be part of the nomological net, the sample from which we collected the data may have not been representative enough to manifest expected relationships. Our sample included teachers who participated in a PLC in the last year, and this may be because they were aware they needed to bolster their CS teaching practices–which could be an indicator of a new CS teacher. Along these lines, most of the teachers included in our sample were relatively new to teaching CS ($M_{years\ teaching\ CS} = 3.90$, $Median_{years\ teaching\ CS} = 3$). Therefore, they may not have had time yet to demonstrate their proficiencies in observable ways (e.g., through CS teaching awards, participation in CSTA activities) which would relate to scores on this measure.

Table 4: Correlations among primary variables

	1	2	3	4	5	6	7	8	9	10	11
1. Total Score											
2. Teaching Experience	0.06										
3. CS Teaching Experience	0.08	0.24*									
4. Teaching Awards	0.19*	0.21*	0.32***								
5. CS Teaching Awards	0.00	0.08	0.06	0.32***							
6. Self-efficacy	0.01	-0.11	0.39**	0.23	0.15						
7. Standards Familiarity	0.07	-0.01	0.21	0.30***	0.16	0.48***					
8. *N* CS Activities	0.02	-0.05	0.14	0.21*	0.19*	0.28*	0.42***				
9. Teach CS Frequency	0.08	0.21	-0.09	-0.05	-0.11	0.03	-0.05	-0.09			
10. Professional Learning Community	-0.25*	-0.02	-0.09	0.15	0.12	0.09	0.09	0.16	-0.15		
11. Other CS PD	-0.02	-0.17	-0.11	0.08	0.04	0.22	0.41***	0.37*	-0.03	0.25	
12. *N* CS PD Completed	0.11	0.05	0.29***	0.24*	0.06	0.36*	0.45***	0.49***	-0.08	0.19	0.54***

*p < .05, **p < .01, ***p < .001.

Table 5: Regression coefficients

Variable	β	$t_{(76)}$	p	95% CI
Teaching Experience	0.15	0.94	> .999	[-0.16, 0.46]
CS Teaching Experience	-0.26	-0.66	> .999	[-1.06, 0.53]
Teaching Awards (Yes)	2.91	0.83	> .999	[-4.05, 9.87]
CS Teaching Awards (Yes)	-0.57	-0.11	> .999	[-11.14, 10.00]
PLC Participation (Yes)	-7.82	-2.70	0.087	[-13.61, -2.04]
Other PD Participation (Yes)	-1.04	-0.29	> .999	[-8.15, 6.07]
N CS Activities	0.13	0.07	> .999	[-3.95, 4.22]
PD Opportunities	1.03	1.18	> .999	[-0.70, 2.76]
Self-assessed Proficiencies	-2.71	-1.00	> .999	[-8.11, 2.68]

Note: 25 participants were excluded due to missing data.
PLC = professional learning community in the last year.

Indeed, we saw evidence of this possibility in the extremely limited variance observed in the number of CS teaching awards received among our participants.

Finally, the measure itself may not be fully capturing teachers' true proficiencies related to the CSTA Teacher Standards. It could be that yet other proficiencies were not assessed, or were under-assessed by our measure. Or, perhaps our measure was not contextualized enough to teachers' own experiences.

We will continue to explore how we can effectively differentiate novice and expert teachers that engage in practices aligned to Standards to further validate the vignettes measure. Unfortunately, many of the ways traditionally used (as defined in 2.3) do not seem practical or viable (e.g., holding licensure at a certain level [10], assessing student learning outcomes [2]) when considering the Standards. We will continue to investigate alternative methods.

5.1 Limitations

While vignettes offer a compromise between self-reflection items and classroom observation, they still do not capture actions that teachers take within their own classroom settings. Therefore, we can only measure their knowledge and decision making within hypothetical classrooms–not their actual planning, assessment, and in-the-moment teaching actions.

The sample may have been heavily skewed towards teachers who knew they needed additional instruction and were more committed to teach CS, given that they took the assessment as part of prework for an upcoming teacher PD. A more representative and larger sample could therefore enable more valid comparisons and generate more variance among all our variables of interest.

Finally, there are other ways to measure a teachers' engagement with Standards, including those presented by Rosato et al. [16]. Although these are still self-reported, the measures used in the study add more nuance and could provide a way to discern between novice and expert teachers in relation to Standards 2, 4, and 5.

6 Conclusion and Future work

The process of creating vignettes to measure teacher proficiencies against the CSTA Standards (2, 4, 5) for CS Teachers has been challenging. We engaged in several formal processes (i.e., cognitive interviews, multiple rounds of development with experts, and pilot testing the vignettes) to bolster the quality of the vignettes assessment. While our attempts to create criteria to distinguish between novice and expert practitioners (of the identified standards) did not provide the insight we had hoped, we remain optimistic that we can revise the vignettes measure and collect additional (and more robust) evidence of validity–including new data collected from teachers participating in future CSTA PD. Nevertheless, this study provides a model for future researchers who wish to develop a vignettes-based measure of teacher proficiencies aligned with a set of teacher standards. This first pilot test rendered mixed results, and thus plenty of work remains to further refine and vet this vignettes-based measure.

Acknowledgments

This project is funded under a grant from the US Department of Education, Education Innovation and Research (EIR) program. Special thanks to Bryan Twarek (Computer Science Teachers Association) for his support and feedback on this project.

References

[1] Stacey Alicea, Katie Buckley, Diana Cordova-Cobo, Aliza Husain, Laura Meili, Lisa Merrill, Krista Morales, Lisa Schmitt, Nathaniel Schwartz, Tim Tasker, Viticia Thames, and Shaye Worthman. 2023. *Measuring Teacher Professional Learning: Why It's Hard and What We Can Do About It.* Technical Report. Research Partnership for Professional Learning (RPPL). 1–24 pages. https://annenberg.brown.edu/sites/default/files/Measuring%20Teacher%20Professional%20Learning.pdf

[2] A. J. Auerbach, M. Higgins, P. Brickman, and T. C. Andrews. 2018. Teacher Knowledge for Active-Learning Instruction: Expert–Novice Comparison Reveals Differences. *CBE—Life Sciences Education* 17, 1 (March 2018), ar12. doi:10.1187/cbe.17-07-0149

[3] Computer Science Teachers Association. [n. d.]. Guidance for Reflective Teachers. https://csteachers.org/reflective-teachers/

[4] Computer Science Teachers Association. 2020. CSTA Standards for CS Teachers. https://csteachers.org/teacherstandards/

[5] Linda Crocker and James Algina. 1986. *Introduction to Classical and Modern Test Theory.* Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887 ($44.75). https://eric.ed.gov/?id=ed312281

[6] Department of Didactics of the Human Sciences, Faculty of Psychology and Educational Sciences, Babeș-Bolyai University, Cluj-Napoca, Romania and Adrian Costache. 2024. The Limitations of the Common Approach and the Educational Value of Teacher Observation. *Educatia 21* 29 (Dec. 2024), 47–54. doi:10.24193/ed21.2024.29.06

[7] Christian Kosel, Elisabeth Bauer, and Tina Seidel. 2024. Where experience makes a difference: teachers' judgment accuracy and diagnostic reasoning regarding student learning characteristics. *Frontiers in Psychology* 15 (March 2024), 1278472. doi:10.3389/fpsyg.2024.1278472

[8] Christian Kosel, Angelina Voggenreiter, Jürgen Pfeffer, and Tina Seidel. 2023. Measuring Teachers' Visual Expertise Using the Gaze Relational Index Based on RealWorld Eyetracking Data and Varying Velocity Thresholds. *Journal of Expertise* 6, 3 (2023), 267–281. https://www.journalofexpertise.org/articles/volume6_issue3/JoE_6_3_Kosel_Voggenreiter_etal.html

[9] Chad Lang and Matt Townsley. 2021. Improving Teacher Evaluation by Walking the Talk of Standards-based Grading: Communicating Educator Growth using Proficiency Scales. *Journal of School Administration Research and Development* 6, 2 (2021), 81–89. https://files.eric.ed.gov/fulltext/EJ1325604.pdf

[10] Yishan Lin, Rui Li, Jesús Ribosa, David Duran, and Binghai Sun. 2024. Expert and Novice Teachers' Cognitive Neural Differences in Understanding Students' Classroom Action Intentions. *Brain Sciences* 14, 11 (Oct. 2024), 1080. doi:10.3390/brainsci14111080

[11] Monica M. McGill, Amanda Bell, Jake Baskin, Anni Reinking, and Monica Sweet. 2023. Measuring Teacher Growth Based on the CSTA K-12 Standards for CS Teachers. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1.* ACM, Toronto ON Canada, 994–1000. doi:10.1145/3545945.3569796

[12] Monica M. McGill, Joseph C. Tise, and Adrienne Decker. 2024. Piloting a Diagnostic Tool to Measure AP CS Principles Teachers' Knowledge Against CSTA Teacher Standard 1. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1.* ACM, Portland OR USA, 819–825. doi:10.1145/3626252.3630905

[13] Daniel Muijs. 2006. Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation* 12, 1 (Feb. 2006), 53–74. doi:10.1080/13803610500392236

[14] Anthony J. Nitko. 1996. *Educational Assessment of Students. Second Edition.* Prentice-Hall Order Processing Center, P. ERIC Number: ED435654.

[15] Ursula Pieper and Jan Vahrenhold. 2020. Critical Incidents in K-12 Computer Science Classrooms - Towards Vignettes for Computer Science Teacher Training. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20).* Association for Computing Machinery, New York, NY, USA, 978–984. doi:10.1145/3328778.3366926

[16] Jennifer Rosato, Joseph Tise, Laycee Thigpen, Fatima Brunson, and Monica McGill. 2025. Coaching as a Means to Support Teacher Development of Computer Science Knowledge and Skills. In *Society for Information Technology & Teacher Education International Conference.* Association for the Advancement of Computing in Education (AACE), 1960–1969. https://csedresearch.org/wp-content/uploads/our_articles/2025CoachingtoSupportTeacherDevelopmentof%20CSKnowledgeandSkills.pdf

[17] Karen Skilling and Gabriel J. Stylianides. 2020. Using vignettes in educational research: a framework for vignette construction. *International Journal of Research & Method in Education* 43, 5 (Oct. 2020), 541–556. doi:10.1080/1743727X.2019.1704243

[18] Laycee Thigpen, Monica M. McGill, Bryan Twarek, and Amanda Bell. 2024. Piloting a Revised Diagnostic Tool for CSTA Standards for CS Teachers. In *Proceedings of the 2024 on ACM Virtual Global Computing Education Conference V. 1.* ACM, Virtual Event NC USA, 207–213. doi:10.1145/3649165.3690122

[19] Charlotte E. Wolff, Halszka Jarodzka, Niek Van Den Bogert, and Henny P. A. Boshuizen. 2016. Teacher vision: expert and novice teachers' perception of problematic classroom management scenes. *Instructional Science* 44, 3 (June 2016), 243–265. doi:10.1007/s11251-016-9367-z

[20] Aman Yadav and Marc Berges. 2019. Computer Science Pedagogical Content Knowledge: Characterizing Teacher Performance. *ACM Transactions on Computing Education* 19, 3 (Sept. 2019), 1–24. doi:10.1145/3303770

[21] David Yeager, Andrew Byrk, Jane Muhich, Hannah Hausman, and Lawrence Morales. 2013. *Practical Measurement.* Technical Report. Carnegie Foundation for the Advancement of Teaching. 1–59 pages. https://www.carnegiefoundation.org/wp-content/uploads/2013/12/Practical_Measurement.pdf