2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #51125

Piloting a Vignettes Assessment to Measure 6th-8th Grade Computer Science Teacher Proficiencies (Fundamental)

Dr. Monica McGill, Institute for Advancing Computing Education

Monica McGill is President & CEO of the Institute for Advancing Computing Education (formerly known as CSEdResearch.org). Having previously worked in industry and academia, McGill uses her experiences as a computer scientist, professor, and researcher to enable others to build a strong foundation of CS and AI education and education research focused on growing access, opportunity, and experiences for learners.

Dr. Joseph C Tise, Institute for Advancing Computing Education

Dr. Tise is a Senior Education Researcher at the Institute for Advancing Computing Education. As an educational psychologist, he has training and expertise in quantitative and mixed-methods education research, particularly within computer science and other STEM domains.

Piloting a Vignettes Assessment to Measure 6th-8th Grade Computer Science Teacher Proficiencies (Fundamental)

Monica McGill[1] and Joseph C. Tise[2]

[1,2]Institute for Advancing Computing Education
[1]monica@csedresearch.org
[2]joe@csedresearch.org

Abstract

Background. The Standards for Computer Science (CS) Teachers by the Computer Science Teachers Association (CSTA) articulate indicators of effective classroom practice. To assist teachers in understanding their proficiency and areas for growth, we designed and pilot-tested a vignette-based assessment of middle school (6th-8th grade) CS teacher proficiencies aligned with Standards 2, 4, and 5.

Research Question. Our research questions for this study was: *How difficult did teachers find the new measure?* and *How did the teachers perform on the vignettes?*

Methodology. We constructed three vignettes with questions aligned to Standards 2, 4, and 5 and geared to middle school teachers and appropriate CS standards. Our cognitive interviews with middle school teachers ($n = 5$) informed iterative revisions of the assessment prior to pilot testing with middle school teachers. We then revised the assessment and piloted it with additional teachers ($n = 26$).

Results. At the scale level, teachers performed moderately ($M = 58.56\%, SD = 15.09\%, Min = 20.00\%, Max = 90.00\%$). The items discriminated fairly well between higher- and lower-performing test takers overall, indicating that performance on each item was generally correlated with performance on all other items in the assessment. The pilot revealed that three items need attention since they were not highly correlated with the performance of the other items.

Implications. This emerging measure of teacher proficiencies is expected to demonstrate promising psychometric properties, though additional revisions to items may be warranted following pilot testing. Once validated, the measure could serve as a useful tool for practitioners to identify areas of strength and opportunities for professional growth, informing future professional development efforts.

1 Introduction and Background

K-12 educators often continually strengthen their content and pedagogical content knowledge—particularly in rapidly evolving fields such as computer science (CS) and, more broadly, engineering. As a subfield of engineering, CS departments are often found in schools and colleges of engineering in higher education due to CS being borne in large part from engineering [1, 2]. As a form of engineering, CS shares many concepts and processes with the broader and more traditional forms of engineering (i.e., mechanical), such as design thinking, problem solving, trouble shooting, abstraction, and testing [3–6]. CS and similar engineering disciplines also shares an interest in K-12 education, including the need to train in-service teachers.

The need for training teachers highlights the importance of reliable assessment tools that can track professional development and support ongoing teacher growth. To support this effort in the field of CS, the Computer Science Teachers Association (CSTA) introduced the Standards for CS Teachers [7], a framework organized around five key areas that support student learning:

1. *CS Knowledge and Content*,
2. *Equity and Inclusion*,
3. *Professional Growth and Identity*,
4. *Instructional Design*, and
5. *Classroom Practice*.

Released in 2019, the Standards reflect the belief that all students are capable of learning CS and engineering-related thinking when supported by teachers who employ inclusive instructional practices and engage in continuous professional learning.

The Standards have since informed PD curricula [8] and guide teacher self-reflection and growth [9] across a variety of self-reported measures. Given that the standards were community-developed with a variety of K-12 teachers, our first attempt was to measure teacher implementation of these standards through their classroom practices [10]. In this previous work, we asked teachers to provide evidence of and justifications for Standards 2, 3, 4, and 5 and created a diagnostic tool for Standard 1 to assess teachers' CS knowledge [10, 11]. However, this evidence-based measure took approximately two hours for teachers to complete; it took even longer to assess with multiple reviewers being trained to use a rubric and assess the evidence.

Given the time commitment needed, we pivoted to develop a multiple-choice, vignettes-based assessment. Vignettes are a standard approach to engage teachers in their pedagogy and practice through carefully constructed scenarios that reflect authentic classroom situations [12–14]. Unlike the evidence-based measure which would be extremely challenging to assess at scale, a multiple-choice assessment can be easily scaled and meaningful–if it has sufficient evidence of reliability and validity. Further, while classroom observation would be the most ideal way to measure teachers' practices, it also has its own limitations [15, 16]. Vignettes can be a practical way to ensure scale can be reached in a practical way and have already been used to measure CS pedagogical content knowledge [13, 17].

After creating and piloting vignettes for K-5 grade teachers who teach CS, we created and piloted vignettes for 6th-8th grade teachers. Our goal was to assess teacher proficiency related to three of the five standards (2, 4, and 5). The indicators for each standard are shown in Table 5. The two

Standards not included in the vignettes included Standard 1, which is focused on CS content knowledge which can be addressed by assessing CS knowledge directly, and Standard 3, which is centered on teacher professional growth which can be self-reported (e.g., *Do you have a CS PD plan?*).

Our research questions for this study were:

RQ1. *How difficult did teachers find the new measure?*
RQ2. *How effectively did each item discriminate between high- and low-performers?*

Table 1: Standards 2, 4, and 5 with their indicators [7].

Standard	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.

2 Methodology

To answer our research questions, we first created a multiple-choice vignette-style assessment that aligned with Standards for CS Teachers 2, 4, and 5. (Standard 1 is focused on content knowledge of CS, and our previous diagnostic [18] assessed this standard. Standard 3 is centered on actions that teachers take to essentially become a more impactful teacher through professional development and was not suitable for classroom scenarios.)

We created the vignettes within the context of the CSTA's 6th-8th grade Curriculum Standards [19]. For example, one of the vignettes centered on *CSTA Standard: 2-AP-12, Design and iteratively develop programs that combine control structures, including nested loops and compound conditionals*. We then created a mini-lesson plan that met this learning standard.

Questions and answer choices on our assessment related to how teachers might plan, teach, and assess that learning standard.

This study included two phases. In the first phase, two researchers and one experienced K-8th grade CS teacher created the assessment. This took multiple rounds of meetings to plan the vignettes and each question stem and response items that were authentic and had one correct answer along with three distractors. Each of the three vignettes had the same set up: a brief description of the students, classroom and school setting followed by a lesson plan that set the stage for the questions and response items. Each question then targeted one indicator from the teacher standards and was grouped into planning, teaching, or assessing.

Each question also had four responses. Our template for creating the four item responses were:

- Ideal action according to rubric/standards: Items which demonstrate clearly good practice (generally and in CS ed) and are 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 (*okay*): Items that won't hurt student learning, but which also likely won't benefit them either.
- Distractor (*bad*): Items that would potentially harm student learning .

We then conducted cognitive interviews with teachers to evaluate the vignettes, the questions, and the responses [20, 21]. After the cognitive interviews were completed, we spent significant time over the course of five weeks reviewing and revising the assessment based on the results of the interviews.

In the follow-up pilot phase, we sent the revised assessment to a larger sample of teachers to generate quantitative data for item-level statistical analyses. The study was approved by an external Institutional Review Board (IRB).

2.1 Participants

For both phases, we recruited 6th-8th grade teachers through the Computer Science Teachers Association (CSTA) membership. CSTA selected teachers who they believed: 1) were proficient in the Standards for Teachers and had significant experience with evidence (e.g., awards received, participation in the CSTA Equity Fellows program); and 2) had unknown levels of experience, particularly targeting newer, less experienced teachers. Teachers who participated in the cognitive interviews received a $90 gift card.

In the pilot testing, we invited 100 teachers to participate and participating teachers received a $30 gift card. Of the teachers who responded ($n = 26$), they averaged nearly 20 years of teaching experience ($SD = 7.48$) and over 10 years teaching CS ($SD = 7.95$). They were moderately familiar with the Teacher Standards 2, 4, and 5 (see Table 3), using a scale from 1 to 7 where 1 was no familiarity and 7 was highly familiar. The standard deviation indicated a wide variety of responses.

Table 2: Years teaching

	Years Teaching	Years Teaching CS
Average	20.08	10.08
Standard Deviation	7.48	7.95
Range	[7-35]	[1-29]

Table 3: Self-reported familiarity with Teacher Standards 2, 4, and 5 (scale from 1 to 7 where 1 was no familiarity and 7 was highly familiar).

	Mean	StDev
Standard 2. Equity and Inclusion	5.62	1.53
Standard 4. Instructional Design	5.65	1.55
Standard 5. Classroom Practice	5.69	1.57

2.2 Data Collection and Analysis Procedures

For the cognitive interviews, we developed a semi-structured interview protocol to guide the interview process. After they read each vignette and item/response, we asked them to choose which they thought was correct and explain why. We then asked them to review each response they did not choose and explain why they did not choose it. We also asked them if the vignette and the item was clear or confusing, whether it contained any bias or deficit-thinking, and whether there were any obviously right or wrong answers. The interview was recorded via Zoom, and the data was analyzed by one researcher who will summarize all feedback from all four interviews into one document. The team of three researchers/teachers then met and discussed all of the responses to each item, revising the assessment based on the interviews.

For the pilot testing quantitative data, we conducted sample descriptive statistics and calculated item-level statistics in line with classical test theory. Specifically, we determined each item's *difficulty* (expressed as a proportion of the sample that answered the item correctly; higher values indicate easier items), and *discrimination* (interpreted as a correlation coefficient, with higher positive values indicating better discriminatory power). Statistically, item discrimination values are point-biserial correlations between the score on an individual item (0 or 1; incorrect or correct) and the total score of the rest of the items on the assessment. Finally, we also collected participants' self-ratings related to each indicator. They rated how likely they were to engage each action related to the indicator (see Table 5) from 1 (Very unlikely) to 5 (Very likely).

3 Results

To answer our research question, we provide a summary of the cognitive interview findings and of the pilot study.

3.1 Cognitive Interview Results

We conducted cognitive interviews with five teachers, two experienced CS teachers and three novices. Two of the novices were in pre-service training programs and one had under five years of experience. Each cognitive interview took place over two sessions in Zoom, and the two sessions lasted approximately two hours in total. The researcher who conducted the interviews took detailed notes and also recorded the interview for further review. All of the key responses from all interviewees were then placed in one document for easier analysis by the research team.

The high-level results from the interviews are presented in Table 4. Although we share the number of correct responses for each item, the cognitive interviews were less about whether the experts outperformed the novices. They were, however, about whether each vignette and question stem and response options were interpreted in the way we had intended. For example, in the item Vignette 1 that targeted indicator 4a, the question stem read:

> As Polly analyzes lesson alignment to the identified standard (CSTA 2-AP-12, Design and iteratively develop programs that combine control structures, including nested loops and compound conditionals), what should her next lesson include to build on this lesson to further meet the standard?

One of the distractors was *Plan a lesson where students create programs that use if/else statements within nested procedures (e.g., functions, methods).* Two of the novices and one of the experienced CS teachers selected this. When considering their responses as to why they chose this, we changed the distractor to: *Plan a lesson where students create programs that use compound conditionals within nested loops.* This distractor would require knowledge about loops and nested loops, so the cognitive load to combine compound conditionals inside nested loops would create a much higher cognitive load.

In another, the question stem read: *Polly knows that 82% of her students in the CS class are boys and several are English language learners. What is the best way she can support inclusion of all students?* One of the distractors was *Polly plans to create diverse groups with all genders and language proficiencies.* One of the expert's responses was *when you've got a majority of class of boys. Sometimes putting the girls together is actually your best bet.* Research has found to support this, and the expert's comments align with this. Our correct answer was *Polly plans to share exemplary programs from girls and English language learners.* However, based on one of the novice's responses: *"exemplary" may be challenging given that [there are] so few girls."* We changed the response to be *Polly plans to share exemplary programs from girls and English language learners from prior years' classes.* This way, there should be exemplary programs that were created by a variety of students.

3.2 Pilot Results

At the scale level, teachers performed moderately ($M = 58.56\%, SD = 15.09\%, Min = 20.00\%, Max = 90.00\%$). The overall pilot results are presented in Table 5. The most difficult item was in Vignette 1, Plan, 2b (1Plan2b); only 12% of our participants answered this item correctly. The easiest item was in Vignette 2, Teach, 5d (2Teach5d) where 85% of the sample answered it correctly. The items discriminated fairly well between higher- and lower-performing test takers overall, indicating that performance on each item was generally correlated with performance on all other items in the assessment. This was true of most items, except items 1Plan4a, 2Plan2c, and 3Plan4b. These three items were not highly correlated with performance on the rest of the assessment, indicating they may be less useful in determining teachers proficiencies related to the standards.

Table 4: Response scores from cognitive interviews. All items were perceived as authentic with no bias.

			Experienced		Novice		Obviously
Vignette	**Target**	**Indicator**	**Correct**	**Incorrect**	**Correct**	**Incorrect**	**Wrong**
1	Plan	**4a.**	0	2	1	2	
1	Plan	**2b.**	1	1	1	2	
1	Plan	**4f.**	1	1	1	2	
1	Teach	**2a.**	1	1	2	1	
1	Teach	**5b.**	0	2	3	0	X
1	Assess	**4f.**	0	2	3	0	
2	Plan	**2c.**	2	0	1	2	X
2	Plan	**4e.**	1	1	3	0	
2	Plan	**4c.**	1	1	3	0	X
2	Plan	**4d.**	2	0	2	1	X
2	Teach	**5d.**	1	1	1	2	X
2	Teach	**5e.**	1	1	1	2	
2	Assess	**5f.**	2	0	2	1	X
3	Plan	**4b.**	0	2	2	1	X
3	Plan	**4a/2e.**	2	0	3	0	X
3	Plan	**5c.**	1	1	3	0	X
3	Teach	**5a.**	2	0	3	0	X
3	Assess	**2d.**	1	1	2	1	X
3	Assess	**2d.**	1	1	1	2	X

4 Discussion

4.1 Cognitive Interviews

None of the interviewees found bias in any of the vignette scenarios, question stems, or responses. All of the interviewees found all of the vignettes, question stems, and responses authentic. This is important, since authenticity is critical in vignettes to ensure that they are representative of a variety of classrooms and that teachers can relate to them [12, 14].

Over half of the items were noted by one or more interviewees as having obviously wrong answers. We investigated each of these, and in about half of the cases we changed the item response and the other half we chose not to. We did not have an issue with an occasional obviously wrong response (as noted by only one or two interviewees) in a multi-item test.

Overall, the interviews were highly informative and gave us confidence that we could move to the pilot stage. For several, we tweaked only minor wording where we learned that interviewees selected certain responses due to issues with semantics. That is, there were cases where the interviewee took the stem or the item response to mean something different than what we had intended.

Interestingly, in some cases (like Indicator 5b in Table 4) the experienced teachers did not choose the correct response and the novices all chose the correct response. This was interesting to us, and as we peered into the reasons why the experts did not choose it, they often projected their

Table 5: Item Statistics for Standards 2, 4, and 5 [7].

Vignette	Target	Indicator	Difficulty	Discrimination	Self-Rated Mean (SD)
1	Plan	**4a**	0.38	0.04	4.04 (0.87)
1	Plan	**2b**	0.12	0.63	4.40 (0.82)
1	Plan	**4f**	0.54	0.33	4.28 (0.84)
1	Teach	**2a**	0.73	0.38	4.31 (0.79)
1	Teach	**5b**	0.42	0.19	4.85 (0.37)
1	Assess	**4f**	0.54	0.47	4.65 (0.49)
2	Plan	**2c**	0.58	0.17	4.31 (0.84)
2	Plan	**4e**	0.81	0.33	4.62 (0.57)
2	Plan	**4c**	0.62	0.58	4.35 (0.80)
2	Plan	**4d**	0.69	0.34	4.81 (0.49)
2	Teach	**5d**	0.85	0.19	4.27 (0.92)
2	Teach	**5e**	0.46	0.35	4.46 (0.99)
2	Assess	**5f**	0.35	0.26	4.62 (0.50)
3	Plan	**4b**	0.77	0.13	4.08 (1.06)
3	Plan	**4a/2e**	0.81	0.30	4.58 (0.64)
3	Plan	**5c**	0.65	0.45	4.54 (0.71)
3	Teach	**5a**	0.77	0.33	4.38 (0.80)
3	Assess	**2d (Access)**	0.38	0.40	4.27 (0.72)
3	Assess	**2d (Participation)**	0.58	0.36	4.27 (0.72)
3	Assess	4g	0.67	0.58	4.19 (1.02)

Note: 2d had multiple parts–one item focused on Access and one focused on Participation.

classroom situation into the vignette rather than referring back to what the vignette teacher would have done that was more appropriate for their classroom scenarios. We also recognized that pre-service teachers in today's teacher education programs have the benefit of learning new methods that may not have been as prevalent two or three decades ago, including that novice teachers may have been taught with more equitable and inclusive practices.

4.2 Pilot Testing

The item-level statistics revealed that our vignette items are generally moderately difficult. Most of the items were answered correctly by close to half of the sample. Additionally, these items were generally moderately correlated with performance on the other items (considered collectively), indicating that each item is providing some accurate information about our participants' proficiencies. Notably, the item statistics reported here indicate somewhat higher-quality items compared to those we have generated in a prior project [22].

We also see in Table 5 that participants tended to indicate they were *Likely* or *Very likely* to engage actions aligned with each indicator of the teacher standards (mean ratings of each were above 4 for all items). This illustrates a potential mismatch between teacher reports of what they would personally do in class, and their actual proficiencies as assessed by our vignettes questions. This potential slight mismatch could indicate that teachers are not accurately judging their behaviors, or that our assessment does not fully capture their true proficiencies, or both.

Additional research to collect more evidence of validity and reliability for the assessment is therefore warranted to ferret out these possibilities.

4.3 Limitations

For the cognitive interviews, an additional experienced teacher would have been beneficial to balance the number of novice teachers. We also do not have a reliable way (yet) to ensure that the experienced teachers are also adhering to standards for teachers–despite our recent undertaking with developing K-5 vignettes [22]. Regarding the pilot testing, a larger and more representative sample would enhance the generalizablity of our findings. As is the case with all item statistics generated within a classical test theory framework, the item statistics are sample dependent and therefore will change as the sample changes.

5 Conclusions and Future Work

Based on the findings from the cognitive interviews and the pilot, we will revise the assessment in an effort to improve its effectiveness. The assessment will be given to over 100 6th-8th grade teachers over the summer as part of their participation in a summer PD week. We will then use the larger set of data to revise the assessment one last time. While we don't intend to place the assessment in the public domain, we will add the assessment to the set of tools that CS PD providers can use to assess teacher understanding of pedagogy, practices, and assessment strategies. Finally, we believe the assessment could be a valuable tool in pre-service teacher preparation programs for in-classroom discussions.

6 Acknowledgments

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

References

[1] Claude E. Shannon. A symbolic analysis of relay and switching circuits. *Electrical Engineering*, 57(12): 713–723, December 1938. ISSN 0095-9197, 2376-7804. doi: 10.1109/EE.1938.6431064. URL `https://ieeexplore.ieee.org/document/6431064/`.

[2] A. Newell, A. Perlis, and H.A. Simon. What is computer science? *Science*, (157):1373–4, 1967. URL `https://www.cs.cmu.edu/~choset/whatiscs.html`.

[3] Noemi V. Mendoza Diaz, Russ Meier, Deborah A. Trytten, and So Yoon Yoon. Computational Thinking Growth During a First-Year Engineering Course. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–7, October 2020. doi: 10.1109/FIE44824.2020.9274250. URL `https://ieeexplore.ieee.org/document/9274250`. ISSN: 2377-634X.

[4] David Jonassen, Johannes Strobel, and Chwee Beng Lee. Everyday Problem Solving in Engineering: Lessons for Engineering Educators. *Journal of Engineering Education*, 95(2):139–151, April 2006. ISSN 10694730. doi: 10.1002/j.2168-9830.2006.tb00885.x. URL `https://onlinelibrary.wiley.com/doi/10.1002/j.2168-9830.2006.tb00885.x`.

[5] Sean Brophy, Stacy Klein, Merredith Portsmore, and Chris Rogers. Advancing Engineering Education in P-12 Classrooms. *Journal of Engineering Education*, 97(3):369–387, July 2008. ISSN 1069-4730, 2168-9830. doi: 10.1002/j.2168-9830.2008.tb00985.x. URL `https://onlinelibrary.wiley.com/doi/10.1002/j.2168-9830.2008.tb00985.x`.

[6] Alannah Oleson, Brett Wortzman, and Amy J. Ko. On the Role of Design in K-12 Computing Education. *ACM Transactions on Computing Education*, 21(1):1–34, March 2021. ISSN 1946-6226. doi: 10.1145/3427594. URL `https://dl.acm.org/doi/10.1145/3427594`.

[7] Computer Science Teachers Association. CSTA Standards for CS Teachers, 2020. URL `https://csteachers.org/teacherstandards/`.

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

[9] Computer Science Teachers Association. Guidance for Reflective Teachers. URL `https://csteachers.org/reflective-teachers/`.

[10] Monica M. McGill, Amanda Bell, Jake Baskin, Anni Reinking, and Monica Sweet. 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*, pages 994–1000, Toronto ON Canada, March 2023. ACM. ISBN 978-1-4503-9431-4. doi: 10.1145/3545945.3569796. URL `https://dl.acm.org/doi/10.1145/3545945.3569796`.

[11] Laycee Thigpen, Monica M. McGill, Bryan Twarek, and Amanda Bell. 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*, pages 207–213, Virtual Event NC USA, December 2024. ACM. ISBN 979-8-4007-0598-4. doi: 10.1145/3649165.3690122. URL `https://dl.acm.org/doi/10.1145/3649165.3690122`.

[12] Carolyn Jeffries and Dale W Maeder. Using Vignettes To Build and Assess Teacher Understanding of Instructional Strategies. *The Professional Educator*, XXVII(1 & 2), 2004. URL `https://files.eric.ed.gov/fulltext/EJ728478.pdf`.

[13] Ursula Pieper and Jan Vahrenhold. 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, pages 978–984, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-6793-6. doi: 10.1145/3328778.3366926. URL `https://dl.acm.org/doi/10.1145/3328778.3366926`.

[14] Karen Skilling and Gabriel J. Stylianides. Using vignettes in educational research: a framework for vignette construction. *International Journal of Research & Method in Education*, 43(5):541–556, October 2020. ISSN 1743-727X, 1743-7288. doi: 10.1080/1743727X.2019.1704243. URL `https://www.tandfonline.com/doi/full/10.1080/1743727X.2019.1704243`.

[15] Department of Didactics of the Human Sciences, Faculty of Psychology and Educational Sciences, Babeș-Bolyai University, Cluj-Napoca, Romania and Adrian Costache. The Limitations of the Common Approach and the Educational Value of Teacher Observation. *Educatia 21*, (29):47–54, December 2024. ISSN 18410456, 22478671. doi: 10.24193/ed21.2024.29.06. URL `http://educatia21.reviste.ubbcluj.ro/data/uploads/article/2024/ed21-no29-art06.pdf`.

[16] Daniel Muijs. Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation*, 12(1):53–74, February 2006. ISSN 1380-3611, 1744-4187. doi: 10.1080/13803610500392236. URL `https://www.tandfonline.com/doi/full/10.1080/13803610500392236`.

[17] Aman Yadav and Marc Berges. Computer Science Pedagogical Content Knowledge: Characterizing Teacher Performance. *ACM Transactions on Computing Education*, 19(3):1–24, September 2019. ISSN 1946-6226. doi: 10.1145/3303770. URL `https://dl.acm.org/doi/10.1145/3303770`.

[18] Monica M. McGill, Joseph C. Tise, and Adrienne Decker. 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*, pages 819–825, Portland OR USA, March 2024. ACM. ISBN 979-8-4007-0423-9. doi: 10.1145/3626252.3630905. URL `https://dl.acm.org/doi/10.1145/3626252.3630905`.

[19] Deborah Seehorn, Tammy Primann, Todd Lash, Bryan Twarek, Daniel Moix, Leticia Batista, Julia Bell, Chris Kuszmaul, Dianne O'Grady-Cunniff, Minsoo Park, Lori Pollock, Meg Ray, Dylan Ryder, Vicky Sedgwick, Grant Smith, and Chimna Uche. CSTA K-12 Computer Science Standards Revised 2017. Technical report, Computer Science Teachers Association, 2017. URL `https://members.csteachers.org/documents/en-us/46916364-83ab-4f51-85fb-06b3b25b417c/1/`.

[20] Heather Howell, Geoffrey Phelps, Andrew J. Croft, David Kirui, and Drew Gitomer. Cognitive Interviews as a Tool for Investigating the Validity of Content Knowledge for Teaching Assessments. *ETS Research Report Series*, 2013(1), June 2013. ISSN 2330-8516, 2330-8516. doi: 10.1002/j.2333-8504.2013.tb02326.x. URL `https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02326.x`.

[21] Christina Hamme Peterson, N. Andrew Peterson, and Kristen Gilmore Powell. Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes. *Measurement and Evaluation in Counseling and Development*, 50(4):217–223, October 2017. ISSN 0748-1756, 1947-6302. doi: 10.1080/07481756.2017.1339564. URL `https://www.tandfonline.com/doi/full/10.1080/07481756.2017.1339564`.

[22] Joseph C. Tise, Monica M. McGill, Vicky Sedgwick, Laycee Thigpen, and Amanda Bell. 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*, pages 1054–1060, St. Louis MO USA, February 2026. ACM. ISBN 979-8-4007-2256-1. doi: 10.1145/3770762.3772604. URL `https://dl.acm.org/doi/10.1145/3770762.3772604`.