Piloting a Diagnostic Tool to Measure AP CS Principles Teachers' Knowledge Against CSTA Teacher Standard 1

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

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

Adrienne Decker
University at Buffalo
Buffalo, New York, USA
adrienne@buffalo.edu

ABSTRACT

Problem. Understanding computer science (CS) teacher CS knowledge primarily relies on self-reported data from participants to understand the learning impact on teachers and improve teacher growth. There is currently a lack of quality instruments to determine where CS teachers need to improve their knowledge.

Research Question. Our research question for this project was: *What are the preliminary psychometric properties of a developed measure for AP CS Principles teachers?*

Methodology. We developed and piloted a diagnostic tool for high school teachers (n = 18) who have been or will be engaged in teaching AP Computer Science Principles (AP CSP). We then administered the diagnostic with two groups of teachers at the CSTA PD week in 2023, and analyzed the results.

Findings. The full 22-item measure demonstrated acceptable reliability (α = .74) in the present sample. However, four items were identified as "low performing" based on low discrimination values. Further, despite the adequate reliability of the full scale, reliability of the individual subscales was lower (0.29-0.54). This may have been caused by low sample size and/or the lower number of items included at the subscale level.

Implications. As a pilot, our analysis of the diagnostic indicated that reliability can be improved by revising certain items. We will use this information to revise these items, then pilot the diagnostic again next year with a larger set of teachers. This diagnostic can then be used by CSTA and the broader CS education community for their repertoire of tools designed to inform professional development topics and practices.

CCS CONCEPTS

• **Social and professional topics** → **Computing education**; **Computing education programs**; **Computer science education**.

KEYWORDS

standards for teachers, CSTA, diagnostic, standard 1, AP CS Principles, AP CSP, professional development

SIGCSE 2024, March 20–23, 2024, Portland, OR, USA

ACM ISBN 979-8-4007-0423-9/24/03…$15.00
https://doi.org/10.1145/3626252.3630905

ACM Reference Format:
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 (SIGCSE 2024), March 20–23, 2024, Portland, OR, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3626252.3630905

1 INTRODUCTION

Professional development (PD) for teachers who teach computer science (CS) is critical for teachers who have limited knowledge and/or pedagogical content knowledge needed to teach CS [28, 36]. Given the fast-changing nature of CS, PD will always be a necessary component of teaching CS similar to how professional computer scientists also need to stay continually up-to-date on technology [11, 17]. PD encourages teachers to routinely reflect on their practices, their unique classroom contexts, and how their practices meet their students' needs [5]. While CS in schools in the United States continues to grow, the types of PD offered to teachers will also grow to meet changing demands. For example, while data science and artificial intelligence were rarely discussed a decade ago, current PD offerings addressing these topics continue to appear. Unfortunately, most PD programs rely on self-reported data to gauge their impact. Further, there is no clear way to assess a teacher's growth over time and to understand the larger landscape of what PD is most needed by teachers.

To address this gap, we developed a set of measures to assess growth in teacher knowledge across the *Standards for CS Teachers* from CSTA [10]. The Standards have been created to identify key knowledge, development and implementation practices in which teachers engage. The purpose of our newly piloted measures is to assess and track teacher progress across these Standards. Beyond this project, a widely accepted measure of teacher growth would enable teachers to identify their areas of need, guide PD programs in supporting their growth, allow schools of education to assess future CS teachers' preparedness, and support policymakers as they develop new endorsement and certification requirements for CS teachers.

The Standards for CS Teachers comprise five Standards [4] as shown in Figure 1:

- Standard 1: CS Knowledge and Content
- Standard 2: Equity and Inclusion
- Standard 3: Professional Growth and Identity
- Standard 4: Instructional Design
- Standard 5: Classroom Practice.

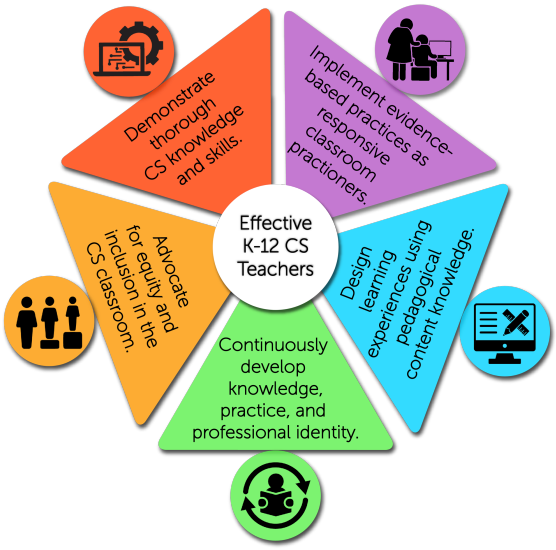


Figure 1: The CSTA Standards for CS Teachers

Each standard consists of five or more indicators with a description of each indicator. Standard 1, CS Knowledge and Content, states "Effective CS teachers demonstrate and continuously develop thorough knowledge of CS content" [4] and has six indicators:

- 1a: Apply CS practices
- 1b: Apply knowledge of computing systems
- 1c: Model networks and the internet
- 1d: Use and analyze data
- 1e: Develop programs and interpret algorithms
- 1f: Analyze impacts of computing

In 2022, we presented a pilot for measuring growth across Standards 2 through 5 [26]. To pilot our new Standard 1 diagnostic tool, we engaged in the following process:

- Develop a diagnostic tool that aligns with Standard 1 and is geared towards Advanced Placement (AP) CS Principles Teachers (high school)
- Pilot the diagnostic with teachers from Indiana
- Collect feedback from participating teachers from Indiana
- Revise the diagnostic based on feedback
- Pilot the revised diagnostic with teachers from South Carolina
- Collect feedback from participating teachers from South Carolina

The primary purpose of this pilot study was to assess preliminary psychometric properties of a newly-developed measure of CS knowledge and skills, aligned with Standard 1 of the CSTA *Standards for CS Teachers*. Our research question was:

> *What are the preliminary psychometric properties of a developed measure for AP CS Principles teachers?*

This study is the first step in developing a diagnostic for a widely-offered CS course that is accessible to beginners and broader than just programming. This study is important for others interested in developing or using tools that provide teachers with knowledge about which areas of growth that they can engage in and PD providers with knowledge needed to make decisions about future PD offerings.

2 BACKGROUND

In this section, we provide a brief background evaluating teacher PD and developing forms of measurement to enable teacher growth.

2.1 Evaluating Teacher PD

Evaluating CS PD across multiple factors is necessary to create and continually revise high-quality, equitable CS PD [2, 12, 19, 27, 37]. Knowledge (content and pedagogical content), skills and beliefs are commonly measured in teacher CS PD [1, 30, 31]. These constructs have been shown to impact student learning and academic growth. For example, teacher content knowledge can map to student learning of the content [6, 21, 24, 33]. The need for teachers to advance their content knowledge (especially in a new subject area) is clear, and we need measures to track such advancement.

The Novice to Expert theory was developed by Benner and Dreyfus, who converted a knowledge development theory for nursing to be used more broadly in education [7, 13]. This theory of the pathway of teachers' learning states that professionals move through five stages of career development (novice, advanced beginner, competent, proficient, and expert, with appropriate rubrics used for each stage), and these stages have an impact on success and career sustainability. As teachers grow in their knowledge, their awareness about where they are in this growth process can enable them to chart their own growth areas for advancement. There is not yet a concept inventory or diagnostic for measuring Standard 1 of the Standards for CS Teachers.

2.2 Forms of Measurement

Measuring teacher PD in any form can be complicated for multiple reasons [28]. The time and resources required to develop a reliable assessment of teacher knowledge gained through PD can be high. More critically, the optics of measuring knowledge can face resistance from teachers. The perception of "testing" teachers can lead to monitoring their test scores rather than using test scores as an indicator to enable professional growth [16, 25]. It can raise questions about *who* may have access to their test scores and *how* their test scores are being used. This is reasonable, since over the decades various ideologies have been promoted to ensure that teachers are held accountable for their students' academic achievement [14, 18].

Previous research has proposed evaluating PD across four key areas: content knowledge, pedagogical content knowledge, self-efficacy/beliefs, and program evaluation [27, 29, 30]. Traditionally, self-reported knowledge gains on surveys have been the primary form used to evaluate teacher CS PD, followed by assessment of content knowledge, and then by interviews [29]. Self-reported measures of understanding of content can be highly susceptible to the Dunning-Kruger effect [15], which is a metacognitive ([34]) phenomenon in which people with low knowledge in an area (like computer science) to rate themselves more highly on their knowledge than they actually are. Having teachers rate themselves on their knowledge in a pre-survey, then rate themselves after they

sat in a PD offering in which they learned their knowledge is actually quite low, can lead to change scores suggesting the teachers experienced learning losses [28].

Assessments (or tests) of content knowledge can mitigate the Dunning-Kruger effect and more accurately assess learning. Still, administering an assessment prior to a participant learning about a subject area can have its own chilling impacts [19], such as on the assessment takers' self-efficacy.

3 METHODOLOGY

To answer our research question, *What are the preliminary psychometric properties of a developed measure for AP CS Principles teachers?*, we first created the diagnostic, then worked with CSTA to identify participants for the pilot study. After data collection, we conducted an analysis of the results.

3.1 Measuring Teacher Knowledge Aligned to Standard 1

To create the diagnostic tool, two researchers with experience in creating assessments for adults engaged in the process of reviewing the CSTA standard 1. One of the researchers was involved in creating the assessment of teachers against their growth areas in Standards 2 through 5, piloting this in the previous year. The other researcher has been involved in creating assessments for the College Board AP exams for CSA over the last decade.

Item design was inspired by the sophisticated and elegant format of publicly-available AP CSP exams provided on the College Board's website [8]. We knew the diagnostic needed to be:

- Accessible to teachers
- Align with Standard 1
- Align with AP CSP courses
- Less than 20 minutes to administer
- Designed to assess content knowledge

Indicator 1a from the standards (Apply CS practices) is an application item. This was not practical for our diagnostic, as ours was to focus on content knowledge, not application. Application would need to be assessed differently and was, in part, covered by the assessment we developed for Standards 2-5. Therefore, we focused on indicators 1b through 1f.

We chose to make the multiple-choice items to ensure ease and consistency in evaluating, to make it more like the AP CSP exam, and to make it quicker to take than having essay or code entry items. Each item included one correct answer and either three or four distractors. Our process for creating the exam involved the following steps:

- Feed the indicators individually into ChatGPT 3.5 and ask it to generate several multiple-choice items for testing
- Take each response and collaboratively reshape it into something meaningful that made sense and accurately reflected the standard
- Provide the assessment for review by two additional people (one with a CS background and one with a K-12 teaching background)
- Revise the assessment based on feedback

The development process took over two days of time for the two researchers. In the end, we created a 22 item multiple-choice measure that assessed knowledge of the indicators as follows:

- 1b: Apply knowledge of computing systems (5 items)
- 1c: Model networks and the internet (4 items)
- 1d: Use and Analyze Data (6 items)
- 1e: Develop programs and interpret algorithms (3 items)
- 1f: Analyze impacts of computing (4 items)

Once completed, the two researchers ranked each item from easiest to most difficult, compared the rankings, then placed all the items in order from easiest to most difficult. The diagnostic was then entered into the REDCap Survey System for further piloting [22, 23].

The researchers also made two additional design decisions. First, the diagnostic was initially designed to only be fully administered if a teacher had taught CS before. If a teacher responded no, then the teacher was exited from the diagnostic. This design decision was intentionally made to help ensure that new CS teachers did not become discouraged at the start of learning about what may be a new subject area for them. Second, the researchers made the decision to provide a range of how the teacher did on the diagnostic rather than provide raw scores (see Figure 2). This decision was similarly made to help ensure even existing CS teachers to not feel discouraged by a raw score of 0.

As an aside, we chose not to provide the items within this paper to protect the integrity of the diagnostic, which is a common practice with these types of instruments [32]. Since it is currently being used internally by CSTA, our intent at the time of this writing is to withhold specific items.

3.2 Participants

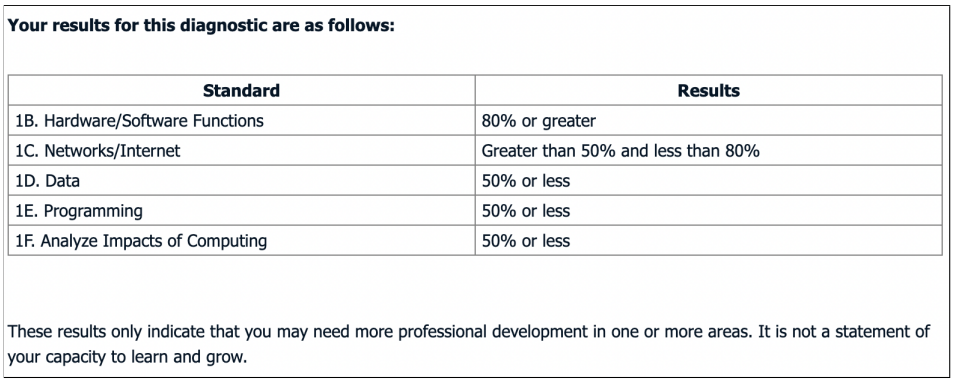
We piloted the diagnostic during two CSTA Professional Development (PD) Weeks in Indiana (June 2023) and South Carolina (July 2023). After discussions with the PD providers, teachers completed the survey on day 1 of the PD as a pre-survey. In the future, the diagnostic may then be used as a post-survey at the end of the school year to help ascertain growth and additional PD needs.

Since we initially set the survey to be taken by current AP CSP teachers and most of the teachers in the PD week in Indiana were new to teaching CS, we only had 6 teachers from Indiana (the first administration of the diagnostic) completing it. Once the diagnostic was taken, we talked to the teachers to get their feedback. The new-to-AP CSP teachers made it clear that they also would like to have taken the diagnostic.

We changed this setting prior to administering the diagnostic in South Carolina, which yielded a higher number of participants. In total, the 18 participants who completed the survey came from 17 high schools across two states: Indiana ($n = 6$) and South Carolina ($n = 12$). The majority were White ($n = 12; 67\%$), Men ($n = 11; 61\%$) and had not yet taught computer science ($n = 11; 61\%$). Six participants (33%) indicated they were Black or African American and seven were Women (39%).

3.3 Analyses

To better understand the quality of the diagnostic, we examined its psychometric properties such as basic descriptive statistics, internal



Your results for this diagnostic are as follows:

Standard	Results
1B. Hardware/Software Functions	80% or greater
1C. Networks/Internet	Greater than 50% and less than 80%
1D. Data	50% or less
1E. Programming	50% or less
1F. Analyze Impacts of Computing	50% or less

These results only indicate that you may need more professional development in one or more areas. It is not a statement of your capacity to learn and grow.

Figure 2: Results were presented in ranges rather than raw scores.

consistency reliability (Cronbach's α; [9]), and item analyses (based in classical test theory, see [35], [20]). Specifically, we report the following for both the full measure and the five individual subscales: internal consistency reliability (Cronbach's α), optimized reliability (i.e., after removing problem items), mean, median, standard deviation, and range. Technically speaking, Cronbach's α is the average correlation of all possible half-test combinations of items (e.g., scores on the first half correlated with scores on the second half, scores on the even items correlated with scores on the odd items). That is, it measures how consistently the items in a test yield similar data from a given respondent. For example, two items designed to assess the same knowledge about programming ought to yield similar scores if the same person answers them. Since each item on a test contributes to the value of α, removing a given item from the test can either improve or diminish α. Thus, part of the analysis process includes recalculating α for remaining items after each given item is removed from analyses. In this way, we can identify items that diminish the quality of the test (i.e., if α improves after removing it).

Additionally 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. 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.

4 RESULTS

4.1 Data Analysis

On average, the sample scored 71.09%(SD = 16.27) on the full diagnostic. The lowest score was 41.00% and the highest was 95.45%.

Table 1 presents descriptive statistics for the full diagnostic and its five subscales. The results indicate that this sample performed best on the *Analyze impacts of computing* and *Model networks and the internet* subscales, and performed worst on the *Use and analyze data* and *Develop programs and interpret algorithms* subscales. The full scale showed acceptable reliability (α = .74), but if we removed four low-performing items (4, 11, 13, & 19) from the diagnostic, the reliability was optimized to α = .81, just above the common rule-of-thumb recommendation of α = .80. These four items are included in subsequent analyses and reporting in this paper. A future iteration will modify or exclude them due to their low discrimination values (i.e., correlations with the other items).

We also examined the scores by state to see if teachers from Indiana vs. South Carolina performed similarly. Indiana teachers (n = 6) averaged 78.79% on the diagnostic, while South Carolina teachers (n = 12) averaged 67.23%. An independent-samples t-test indicated this difference was not statistically significant ($t_{(16)}$ = 1.47, p = .16).

The full 22-item measure demonstrated acceptable reliability (α = .74) in the present sample. However, four items were identified as "low performing" (Items 4, 11, 13, & 19) based on low discrimination values (see Tables 2 and 3). Low discrimination values (operationalized as < .10) indicate that the given item cannot reliably discriminate between participants who scored higher on the rest of the diagnostic compared to those who scored lower. Thus, such items are not accomplishing the diagnostic's objective (to differentiate higher from lower knowledge participants) and ought to be reworded or removed. Removing these four items improved the full-scale reliability to .81.

Despite the adequate reliability of the full scale, reliability of the individual subscales was lower (Table 1). The lower reliability at the subscale level was expected given how Cronbach's alpha is calculated (see [9]). Alpha is partially dependent on the number of items and the observed variances of those items included for analyses. Since each subscale included significantly fewer items than the full-scale, and since each item was scored dichotomously (i.e., 0 or 1), the range of possible observed variances of each item was limited. Such limited range in variance likely negatively impacted the subscale reliability estimates.

4.2 Feedback from Participating Teachers

During our discussions with teachers, only 4 of the teachers from Indiana and no teachers from South Carolina were familiar with

Table 1: Descriptive statistics for the full scale and subscales. Optimized α is the scale α with problem items removed.

	N items	Mean (SD)	Min - Max	α (optimized α)
Full scale	22	71.09 (16.27)	41 – 95.45	.74 (.81)
Use and analyze data	6	57.41 (26.34)	0 – 100	.51 (.56)
Apply knowledge of computing systems	5	76.11 (19.75)	40 – 100	.37 (.43)
Analyze impacts of computing	4	81.94 (20.66)	25 – 100	.29 (.70)
Model networks and the internet	4	83.33 (14.85)	50 – 100	N/A* (.39)
Develop programs and interpret algorithms	3	59.26 (37.15)	0 – 100	.54 (.73)

Note: Descriptive statistics are expressed in percentages
*α not calculable because of zero variance in item 17, & very low discrimination on item 19. Optimized α calculated from only items 16 & 18

Table 2: Item statistics. Item numbers do not reflect the number of the item as presented in the diagnostic.

Item	Difficulty	Discrim.	α if deleted	Subscale
Item 1	0.94	0.44	0.73	Data
Item 2	0.56	0.37	0.71	Data
Item 3	0.65	0.49	0.71	Data
Item 4	0.29	0.01	0.76	Data
Item 5	0.59	0.34	0.73	Data
Item 6	0.53	0.25	0.74	Data
Item 7	0.97	0.29	0.74	Function
Item 8	0.72	0.51	0.71	Function
Item 9	1.00	-	-	Function
Item 10	0.61	0.28	0.74	Function
Item 11	0.53	0.13	0.75	Function
Item 12	0.72	0.33	0.73	Impacts
Item 13	0.67	0.05	0.76	Impacts
Item 14	0.89	0.55	0.72	Impacts
Item 15	1.00	-	-	Impacts
Item 16	0.94	0.15	0.74	Networks
Item 17	1.00	-	-	Networks
Item 18	0.83	0.68	0.70	Networks
Item 19	0.59	-0.02	0.76	Networks
Item 20	0.76	0.56	0.71	Programming
Item 21	0.53	0.37	0.72	Programming
Item 22	0.59	0.46	0.72	Programming

Standard 1 from the CSTA Standards for Teachers. This is not surprising, given that the number of teachers who had not taught CS was very high and, for many of them, this was the first opportunity to learn about teaching CS.

For those in Indiana who selected that they did not have any experience teaching CS and subsequently exited from the diagnostic before taking it, many remarked that they would have liked to take it to understand where they needed to grow. As one participant said, "I think it would've been helpful to see where I fell in my own knowledge." Another remarked that "I do not think it would have been harmful or helpful as I entered with very novice abilities and levels of knowledge." These remarks from the Indiana teachers convinced us to ultimately allow South Carolina teachers with no CS teaching experience to also take the diagnostic.

Regarding the score range provided to them (rather than the raw score), several remarked that the range was unnecessary. One participant remarked that "We are all professionals, and we can handle a low score. It may motivate us to know the level of learning that we need to face to be prepared for the classroom."

As far as the time it took to complete the diagnostic, all participants completed it under 15 minutes, with some taking much less (as captured by one of the researchers present when the diagnostic was administered). This data will be captured in future administrations of the diagnostic.

5 DISCUSSION

5.1 General

Overall, the newly-created diagnostic shows promising psychometric properties and, with some modifications, could be used reliably for both research and practice. In addition to modifying or removing the four problem items identified, we found that some items could still benefit from further refinement. For instance, several items included three individual distractor items (i.e., options A, B, C) and a fourth distractor option "All of the above" (option D).

A small subset of these items included option D (All the above) as the correct answer. It is best not to include "All of the above" as the correct answer, because a participant may feasibly only know that option A & C are correct but does not know option B is also correct.

Table 3: Items with poor answer structure; each item has an "All of the above" option.

Item	Responses
What is the purpose of creating a computational model?	• To predict future technological advancements of block-chain applications • To analyze and understand complex systems • To generate data that can be used for machine learning • To predict the output of an algorithm
How can social media influence socio-political discourse?	• By facilitating the spread of misinformation and propaganda • By promoting critical thinking and informed decision-making • By encouraging respectful and constructive dialogue • By promoting equal representation of diverse perspectives
A company is deciding whether to use biometric authentication or password authentication for its employees to access company data. Which factor should the company consider in making this decision?	• The cost of implementing each authentication method. • The length of time each employee has worked for the company. • The type of publicly-available data being accessed. • The length of time each employee spends accessing the data.
What is an example of a typical, first set of troubleshooting strategies for technology when helping students in a lab?	• Ensuring hardware and software compatibility, resolving connectivity problems, updating the driver, and configuring the system settings. • Restarting the device, checking the power supply, replacing all hardware components, and checking the network connectivity. • Conducting a factory reset, reinstalling the software, performing a virus scan, and adjusting the screen resolution. • Removing unnecessary software, uninstalling applications, and clearing the trash.

However, this participant may still get the item correct based only on their knowledge of options A & C and their test-savviness to recognize that, if two options are correct, option D (All the above) must be the correct choice–despite knowing nothing about option B. We would have no way to differentiate their knowledge from a different participant who knew options A, B, & C were all correct.

Finally, we observed that three items (9, 15, & 17) had a difficulty score of 1.00. This indicates that every participant answered the item correctly, and thus these items contribute no unique information to the diagnostic. While it is beneficial to include a small number of easy items (in which all or nearly all participants answer them correctly) to promote motivation, researchers must balance this against the limited time allocated for a given diagnostic.

5.2 Limitations

The data presented in this study were collected during one data collection instance, and the diagnostic was created specifically for teachers who teach AP CSP at the high school level. The sample was also limited to a small, relatively demographically homogeneous group of high school teachers. Further, the majority of the sample were new to the subject area; the scale- and item-level statistics may differ for samples comprised mostly of experienced or more-knowledgeable teachers. Still, the descriptive statistics for the full diagnostic showed adequate variance, and given the average performance, no floor or ceiling effects were observed. Finally, some types of knowledge are much more difficult to assess via multiple-choice (e.g., ability to critique and/or generate information; see [3]), and thus this diagnostic cannot be considered a panacea for all needs.

6 CONCLUSION AND FUTURE WORK

The CSTA Standard 1 for Teachers diagnostic that we piloted in this study provided very specific and critical information for improving it. Using this data, we will take the next two years of the project to revise the items based on the results of this analysis, using classical theory test to provide insight into how the diagnostic changes over time. We will also be opening the diagnostic up to a larger number of participants prior to summer 2024 for an additional test. This diagnostic can then be used by CSTA for their repertoire of tools designed to inform professional development topics and practices.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Daniel Alston, Jeff. Marshall, and Andrew Tyminski. 2017. Convincing Science Teachers for Inquiry-Based Instruction: Guskey's Staff Development Model Revisited. *Science Educator* 25 (2017). Issue 2.

[2] Judy Anderson and Deborah Tully. 2020. *Designing and Evaluating an Integrated STEM Professional Development Program for Secondary and Primary School Teachers in Australia.* https://doi.org/10.1007/978-3-030-52229-2_22

[3] Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2001. A Taxonomy for Teaching, Learning, and Assessment.

[4] Computer Science Teachers Association. [n. d.]. Standards for CS Teadchers. https://www.csteachers.org/page/standards-for-cs-teachers. Accessed: 2022-08-19.

[5] Beatrice Avalos. 2011. Teacher professional development in Teaching and Teacher Education over ten years. *Teaching and Teacher Education* 27, 1 (2011), 10–20. https://doi.org/10.1016/j.tate.2010.08.007

[6] Jürgen Baumert and Mareike Kunter. 2013. The effect of content knowledge and pedagogical content knowledge on instructional quality and student achievement. In *Cognitive activation in the mathematics classroom and professional competence of teachers*. Springer, 175–205.

[7] Patricia Benner. 1982. From novice to expert. *American Journal of nursing* 82, 3 (1982), 402–407.

[8] College Board. 2023. AP Computer Science Principles. https://apcentral.collegeboard.org/courses/ap-computer-science-principles/exam

[9] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[10] CSTA. [n. d.]. Standards for CS Teachers. https://csteachers.org/teacherstandards/ [Accessed: (15 August 2023)].

[11] Linda Darling-Hammond and John Bransford (Eds.). 2007. *Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do*. Jossey-Bass, San Francisco, CA, USA.

[12] Sloan Davis, Jason Ravitz, and Juliane Blazevski. 2018. Evaluating Computer Science Professional Development Models and Educator Outcomes to Ensure Equity. *2018 Research on Equity and Sustained Participation in Engineering, Computing, and Technology, RESPECT 2018 - Conference Proceedings*. https://doi.org/10.1109/RESPECT.2018.8491716

[13] Stuart E Dreyfus. 2004. The five-stage model of adult skill acquisition. *Bulletin of science, technology & society* 24, 3 (2004), 177–181.

[14] Michael A Dubrovich. 2002. Student Achievement Data: Holding Teachers Accountable. *Principal* 81, 4 (2002), 30.

[15] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*. Vol. 44. Elsevier, 247–296.

[16] Eyvind Elstad, Eli Lejonberg, and Knut-Andreas Christophersen. 2015. Teaching evaluation as a contested practice: Teacher resistance to teaching evaluation schemes in Norway. *Education Inquiry* 6, 4 (2015), 27850.

[17] Michelle Friend, Bryan Twarek, James Koontz, Amanda Bell, and Abigail Joseph. 2022. Trends in CS Teacher Professional Development: A Report from the CSTA PD Committee. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1* (Providence, RI, USA) *(SIGCSE 2022)*. Association for Computing Machinery, New York, NY, USA, 390–396. https://doi.org/10.1145/3478431.3499292

[18] Jack Frymier. 1998. Accountability and student learning. *Journal of Personnel Evaluation in Education* 12 (1998), 233–235.

[19] Thomas R Guskey. 2003. What makes professional development effective? *Phi delta kappan* 84, 10 (2003), 748–750.

[20] Ronald K. Hambleton and Russell W. Jones. 1993. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice* (1993), 38–47.

[21] Karla Hamlen, Nigamanth Sridhar, Lisa Bievenue, Debbie K. Jackson, and Anil Lalwani. 2018. Effects of teacher training in a computer science principles curriculum on teacher and student skills, confidence, and beliefs. *SIGCSE 2018 - Proceedings of the 49th ACM Technical Symposium on Computer Science Education* 2018-January. https://doi.org/10.1145/3159450.3159496

[22] Paul A Harris, Robert Taylor, Brenda L Minor, Veida Elliott, Michelle Fernandez, Lindsay O'Neal, Laura McLeod, Giovanni Delacqua, Francesco Delacqua, Jacqueline Kirby, et al. 2019. The REDCap consortium: building an international community of software platform partners. *Journal of biomedical informatics* 95 (2019), 103208.

[23] Paul A Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G Conde. 2009. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics* 42, 2 (2009), 377–381.

[24] Melanie M Keller, Knut Neumann, and Hans E Fischer. 2017. The impact of physics teachers' pedagogical content knowledge and motivation on students' achievement and interest. *Journal of Research in Science Teaching* 54, 5 (2017), 586–614.

[25] Jihyun Kim and Peter Youngs. 2016. Promoting instructional improvement or resistance? A comparative study of teachers' perceptions of teacher evaluation policy in Korea and the USA. *Compare: A Journal of Comparative and International Education* 46, 5 (2016), 723–744.

[26] 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*. 994–1000.

[27] Monica M McGill, Leigh Ann DeLyser, Karen Brennan, Baker Franke, Errol Kaylor, Eric Mayhew, Kelly Mills, and Aman Yadav. 2020. Evaluation and assessment for improving CS teacher effectiveness. *ACM Inroads* 11, 4 (2020), 35–41.

[28] Monica M McGill, Rebecca Zarch, Stacey Sexton, Julie M Smith, Christine Ong, Melissa Rasberry, and Shelly Hollis. 2021. Evaluating computer science professional development for teachers in the united states. In *Proceedings of the 21st Koli Calling International Conference on Computing Education Research*. 1–9.

[29] Muhsin Menekse. 2015. Computer science teacher professional development in the United States: a review of studies published between 2004 and 2014. *Computer Science Education* 25, 4 (2015), 325–350.

[30] Emmelien Merchie, Melissa Tuytens, Geert Devos, and Ruben Vanderlinde. 2018. Evaluating teachers' professional development initiatives: towards an extended evaluative framework. *Research Papers in Education* 33 (2018). Issue 2. https://doi.org/10.1080/02671522.2016.1271003

[31] Shivaun O'Brien, Gerry McNamara, Joe O'Hara, and Martin Brown. 2020. Learning by doing: evaluating the key features of a professional development intervention for teachers in data-use, as part of whole school self-evaluation process. *Professional Development in Education* (2020). https://doi.org/10.1080/19415257.2020.1720778

[32] Miranda C Parker, Mark Guzdial, and Shelly Engleman. 2016. Replication, validation, and use of a language independent CS1 knowledge assessment. In *Proceedings of the 2016 ACM conference on international computing education research*. 93–101.

[33] Jason Ravitz, Chris Stephenson, Karen Parker, and Juliane Blazevski. 2017. Early lessons from evaluation of computer science teacher professional development in Google's CS4HS program. *ACM Transactions on Computing Education (TOCE)* 17, 4 (2017), 1–16.

[34] Gregory Schraw and David Moshman. 1995. Metacognitive theories. *Educational Psychology Review* 7 (1995), 351–371. Issue 4. https://doi.org/10.1007/BF02212307

[35] Ross E. Traub. 1997. Classical Test Theory in historical perspective. *Educational Measurement: Issues and Practice* (1997), 8–14.

[36] Aman Yadav, Alex Lishinski, and Phil Sands. 2021. Self-efficacy Profiles for Computer Science Teachers. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 302–308.

[37] Albert Zeggelaar, Marjan Vermeulen, and Wim Jochems. 2020. Evaluating effective professional development. *Professional Development in Education* (2020). https://doi.org/10.1080/19415257.2020.1744686