Developing the Physics Teacher Education Program Analysis rubric: Measuring features of thriving programs

Stephanie V. Chasteen*
Chasteen Educational Consulting, UBC 390, Boulder CO 80309, USA

Rachel E. Scherr
University of Washington, Bothell, Washington 98011, USA

(Received 26 August 2019; accepted 13 January 2020; published 3 April 2020)

Given the insufficient number of well-qualified future physics teachers in the U.S., physics programs often seek guidance for how to address this national need. Measurement tools can provide such guidance, by both defining excellence in physics teacher education (PTE) and providing a means to measure progress towards excellence. This paper describes the development of such a measurement tool—the Physics Teacher Education Program Analysis rubric. The rubric was developed by identifying common features and practices at 8 "thriving" PTE programs, defined as large U.S. programs consistently graduating 5 or more future physics teachers in a year. The rubric consists of 89 items, each with 3 levels of achievement (developing, benchmark, exemplary), plus a not present level, which are organized into 6 standards. The rubric has demonstrated a variety of forms of validity, including a strong theoretical basis, empirical validation through program visits, and expert review. The rubric and its associated supporting materials are intended to help program leaders in using a process of continuous improvement and assessment to strengthen existing PTE programs or to establish new pathways for student licensure. The rubric also provides substantive opportunities for research, through further validation and development of the rubric, and by using rubric results to learn more about effective practices in physics teacher education.

DOI: 10.1103/PhysRevPhysEducRes.16.010115

I. INTRODUCTION

The lack of well-qualified physics teachers in the U.S. has ongoing negative repercussions for physics instruction and the number of students entering STEM careers in this country [1,2]; an influx of new physics teachers is needed to address this challenge. However, our country has yet to establish numerous effective programs to prepare future physics teachers; the National Task Force on Teacher Education in Physics (TTEP) task force found that "nationally, physics teacher preparation is inefficient, incoherent, and unprepared to deal with the current and future needs of the nation's students." [1]. We present one tool intended to address this programmatic need: a rubric that allows measurement of what thriving physics teacher education (PTE) programs do and gives specific guidance to institutions wishing to improve their PTE program.

*stephanie@chasteenconsulting.com

A. The need for strong PTE programs

To address the national need for qualified physics teachers, three main types of changes to PTE programs are needed:

1. Increase the number of institutions with PTE programs,
2. Increase the number of graduates from existing PTE programs, and
3. Improve the quality of preparation within existing PTE programs.

To the first point, few U.S. institutions prepare physics teachers: At the time of TTEP, 43% of U.S. institutions had no PTE programs (within either the physics department or school or college of education) and most physics departments graduate no future physics teachers in a two-year period [1]. Only about 20% of physics departments have a track for future physics teachers that has graduated even one teacher [3]. Even among departments with PTE programs, very few license more than two future physics teachers in a year. Meeting the national need will require both creating more PTE programs and increasing the number of graduates from existing programs. However, quality is as important as quantity; despite substantial evidence that effective teachers require both preparation in disciplinary content *and* pedagogy [4–8], at least half of physics teachers do not have a major in the discipline [1,9]

and few learn physics content using the inquiry techniques that are important both for solidifying concepts and for modeling effective physics instruction [10,11] Additionally, most PTE programs lack physics-specific pedagogical preparation [1] with at most a single course in physics pedagogy, and more often general science pedagogy or methods courses.

The Physics Teacher Education Coalition (PhysTEC) was established in 2001 to improve and promote the education of future physics teachers through a wide variety of activities, including funding programs, undertaking research, and supporting the development of a PTE community [12]. PhysTEC has made substantial progress towards improving PTE nationwide—engaging departments around the country, and tripling the number of physics teachers prepared at supported sites [13], but there is more to be done. The current work is intended to build upon and extend these prior successes.

B. The goal: Continuous improvement

The factors underlying the above-named challenges in PTE are complex and systemic, and related to the myriad challenges in revamping undergraduate education in physics [14–16] and other fields [17,18]. Previous reports have made a variety of recommendations to undergraduate programs, all of which follow a model of *continuous improvement*, using a cyclical action-research approach of "Plan-Act-Observe-Reflect" [19]: Establish goals and vision, identify gaps and opportunities, choose strategies to address gaps (based on the best available evidence), implement changes, assess the results, and use those results to inform future program modifications. This general process applies for a wide variety of goals (e.g., to improve student learning within a course, to prepare students for diverse careers, or to recruit more students to the major) and the improvements might include course design, redesign of the curriculum or major, or other activities. These reports also recommend fostering some of the habits and attitudes that support systemic change within a complex system: a community of practitioners, engaged leadership, and a culture of continuous improvement and experimentation among those enacting the change.

In sum, creating change in undergraduate programs requires a goal-driven process of continuous assessment and improvement, guided by energetic leaders in a supportive and knowledgeable community. A current example of this approach is modeled in the Effective Practices for Physics Programs (EP3) project (a joint project of APS and AAPT [20]) which is generating a guide for self-assessment of physics programs, to assist programs in developing such a culture of continuous self-improvement. Local leaders in physics and education wishing to improve the number of qualified physics teachers they produce ought to enact a similar process. In particular, they should

1. Analyze their local landscape
2. Choose strategies that will help address their local challenges
3. Assess the results of those changes, and
4. Use those results to guide future improvements.

A wealth of knowledge has been accumulated in recent years as to the most effective practices for recruiting and educating future physics teachers [1,21,22]. Both PhysTEC and TTEP have identified features that seem important for physics teacher preparation, such as a dedicated faculty champion, an institutional mission that supports teacher preparation, and physics-specific pedagogy courses, and have illustrated those features with many rich examples from institutions. Additionally, PhysTEC has invested substantial resources in creating a community of scholars and practitioners in PTE (e.g., through annual conferences, member institution status, recognition for local achievements, and other mechanisms). This list of PTE features and community of scholars and practitioners provide a strong base to support local leaders in continuous improvement of their PTE programs—but more is needed: a specific tool to guide PTE programs in achieving excellence.

C. Why a rubric?

While the lists of PTE program features offer some helpful guidance, "If you can't measure it, you can't improve it." [23]. The philosophy of the current project is that a measurement tool will support growth in PTE programs by (a) cataloging what PTE programs should do, (b) allowing measurement and thus feedback on achievement of these practices, and (c) supporting reflection and continuous improvement.

Rubric development is a particularly valuable opportunity to surface implicit values and desired outcomes, generating shared understanding about what matters [24,25]. Because a rubric allows different levels of achievement (rather than only a yes or no checkbox), it is a structure for measuring not just the *presence* of an attribute but the *quality* of that attribute [26]. Rubrics can be valuable tools even in complex situations [25–27]. Rubrics are commonly used tools in program evaluation and accreditation processes. Requiring a program leader to commit to rating their program at a specific level on each item likely supports critical thinking about the current state of their program over time.

This paper will present this measurement tool, the Physics Teacher Education Program Analysis (PTEPA) rubric, with the aim of establishing the rubric as a valid and useful tool for the physics education community. In what follows we will

- Describe the development, structure, and validity testing of the rubric, to establish the trustworthiness of the instrument.
- Illustrate how program leaders can use the instrument and what they might learn, to encourage them to make use of the instrument.

TABLE I. The PTEPA rubric. This shows standards, components, and example items from PTEPA rubric version 2.0 [30]. Alphanumeric code preceding each component (e.g., "1A") is the designation of the component number.

Standards and Components	Number of items	Example items
Standard 1: Institutional commitment	14	
There is a strong institutional commitment to STEM teacher education, supported by policy, rewards, and financial resources.		
1A: Institutional climate and support		1A-4 University-level support for teacher education
1B: Reward structure		1B-2 Time for PTE program leaders to engage
1C: Resources		1C-1 Engaged staff
Standard 2: Leadership and collaboration	21	
The program has an effective leadership team, including effective collaboration between physics and education.		
2A: Program team members		2A-1 PTE program leaders
2B: Program team attributes		2B-2 Positional power
2C: Program collaboration		2C-1 Communication across units on PTE program elements
Standard 3: Recruitment	19	
The program recruits many physics teacher candidates by taking advantage of local opportunities and offering attractive options for participation.		
3A: Recruitment opportunities		3A-1 Physics majors
3B: Recruitment activities		3B-2 Physics teaching ambassador
3C: Early teaching experiences for recruiting teacher candidates		3C-2 Exposure to intellectual challenge of teaching
3D: Streamlined and accessible program options		3D-1 Undergraduate licensure pathway
Standard 4: Knowledge and skills for teaching physics	12	
The program ensures that teacher candidates are well prepared to teach physics effectively through rigorous and experiential preparation in physics content and pedagogy.		
4A: Physics content knowledge		4A-1 Physics degree for teacher candidates
4B: Pedagogy courses and curriculum		4B-3 Disciplinary content of certification coursework
4C: Practical K-12 school experiences		4C-3 Field experiences in physics
Standard 5: Mentoring, community, and professional support	11	
The program provides mentoring and induction to support progress toward degree, certification, and retention in the profession, supported by strong student community.		
5A: Mentoring and community support toward a physics degree		5A-1 Student community in physics
5B: Mentoring and community support toward becoming a physics teacher		5B-2 PTE mentor for physics teacher candidates
5C: In-service mentoring and professional community		5C-4 Professional development for in- service teachers
Standard 6: Program assessment	12	
The program assesses multiple outcomes, using them for program improvement and to advocate for funding and resources.		
6A: Program outcomes		6A-2 Annual recruitment in PTE program
6B: Program evaluation and improvement		6B-2 Feedback from stakeholders
6C: Communication to stakeholders		6C-2 Communication with university administrators

A companion paper [28] presents what has been learned so far from scholarly application of the rubric.

II. THE PTEPA RUBRIC

To better orient the reader as to the discussion that follows, we will first summarize the PTEPA rubric; greater detail on its development and properties are given later in this paper. The PTEPA rubric was developed to describe activities and structures observed at 8 thriving PTE programs. "Thriving" programs are defined as programs at large institutions that consistently graduate 5 or more highly qualified future physics teachers each year. The choice to use a threshold of 5 teachers is based on the fact that very few (about 1%) physics programs graduate as many as 5 teachers per year; only about 1%, based on both a nationwide survey in 2012 (6 out of 578 departments) [1] and more recently available Title II data (19 out of 1584 programs) [29]. Most physics departments (84%) graduate no physics teachers [29]. Thus, an increase of even one teacher per year by a department is a significant achievement towards meeting the national need, even if it does not close the gap completely. These thriving programs are those which—compared to others in the U.S.—are doing something different, and thus worth investigating systematically.

The rubric catalogs the features observed at these 8 thriving programs, and the range of performance that could

be associated with those features. The resulting rubric consists of 89 items, each with three levels: "developing", "benchmark", and "exemplary" (plus "not present"), with benchmark representing a recommended level of achievement. Some of the individual items are identified as "prevalent," meaning that thriving programs consistently demonstrated strength in that item.

The items are organized into six "standards": (1) Institutional commitment, (2) Leadership and collaboration, (3) Recruitment, (4) Knowledge and skills for teaching physics, (5) Mentoring, community, and professional support, and (6) Program assessment. Each standard has three or four "components" within it that address specific subtopics. The overall structure of the rubric is shown in Table I. The version of the rubric as of this writing is version 2.0 (released August 2018); the latest version can be obtained at the website in Ref. [30].

The rubric was developed specifically to characterize features of *physics* teacher education programs, and so focuses on the disciplinary preparation of future physics teachers (rather than including items that lie primarily in the domain of a college or school of education or are features of science teacher education in general). The rubric is intended to support self-assessment and improvement of PTE programs, rather than to provide a checklist of universal standards that all programs must follow.

Because this is an instrument intended to reflect practices and structures in PTE in the United States, it will necessarily diverge somewhat from the recommendations which might be made for other countries. For example, outside the U.S. it is more common to have PTE be led by professionals in physics teacher education and to be led from colleges or schools of education (or equivalent) [31]. Thus, the rubric's focus on developing strong leadership from the institution and within the physics department (standards 1 and 2) likely reflects the nature of PTE in the U.S. The experience of in-service teachers will also vary widely across countries, including salary, prestige, and mentoring opportunities—likely impacting the required recruiting practices (standard 3) and in-service mentoring requirements (component 5C), for example. However, programs in the U.S. also bear many commonalities across programs, including the importance of training in physics content, pedagogy, and nature of science, field work, and knowledgeable mentors [31,32]. Indeed, in a pilot application of the rubric to a single non-U.S. program, we noted similar performance across the rubric, with the notable exception of areas of recruiting and mentoring (components 3B, 3C, and 5C), which were rated lower than the programs examined for this study.

III. RUBRIC DEVELOPMENT

The PTEPA rubric was developed by examination of existing literature and instruments to develop a draft rubric and then applying that draft rubric in program visits to thriving programs. Again, thriving programs are programs at large institutions that consistently graduate 5 or more highly qualified future physics teachers each year. The rubric was iteratively improved to reflect observations at the program visits, reconciled with the literature and expert knowledge, and organized to enhance usability. An overview of the year-and-a-half-long development of the rubric is provided in Ref. [33].

A. Initial draft

We based the initial draft of the PTEPA on an existing instrument, the Teacher Education Program Assessment (TEPA) [27,34], which was developed through the Science and Mathematics Teacher Imperative to provide a better understanding and assessment of practices and design features of high-quality teacher preparation. Several items and sections of the TEPA were removed from our draft because they were not specific to the domain of the PTEPA rubric (which focused on *physics* teacher education), or modified to better reflect the known features of quality PTE programs as embodied in the PhysTEC Key Components [35]. The TEPA had a 4-point rating scale, but this scale was *generic* (there was no description of what each level of achievement might look like). For the PTEPA rubric, we developed instead an *analytic* rubric, in which each level of achievement ("scale point") is described specifically [25,26]. This choice was intended to enhance the reliability of ratings and provide more concrete descriptions of achievement in each area on the rubric. The initial draft rubric included 36 items, each with 5 scale points tentatively labeled absent, benchmark, milestone 1, milestone 2, and strength.

B. Program visits

We intended the PTEPA rubric to reflect empirical results of what was observed at thriving programs, rather than to be a theoretical description of *a priori* "best practices." A total of 8 thriving programs were chosen for this study. All were programs at large universities that typically graduate five or more highly qualified physics teachers in a year and thus belong to the "5+ Club," an award provided by PhysTEC to PTE programs graduating 5 or more future physics teachers.

The final list of programs was chosen for their diversity: Half had received PhysTEC funding in the past and half had not, and they had a wide variety of structures. Inclusion of programs that had not received PhysTEC funding was important for avoiding circular logic (since PhysTEC-funded programs are required to address the Key Components, which were included by design in the development of the PTEPA rubric). The list of programs visited is shown below; an asterisk indicates programs that had received PhysTEC funding.

- University of Texas at Austin
- University of Colorado Boulder*

- Brigham Young University
- California Polytechnic University San Luis Obispo*
- Georgia State University*
- Rowan University*
- Rutgers University
- Stony Brook University

For each program, one of us (S. V. C. or R. E. S.) conducted a series of interviews with a variety of program staff over a two-day program visit (in person or virtual). (The report *A Study of Thriving Physics Teacher Programs* [36] offers more detail on program visits, including interview protocols.) We used the results from the interviews to rate each program on the draft rubric, with detailed notes justifying the rating. We shared these ratings with the program leaders for feedback. In some cases, the program leader's feedback conflicted with other information we gathered during the program visits; in such instances, we did not necessarily default to the program leader's rating but instead used the best available data to determine the rating (to avoid potential bias from the program leader).

After each program visit, we modified the PTEPA rubric items, sometimes adding new items or modifying the wording of items or scale points. Items were not typically *removed* until later in the process, once aggregate results were available across all programs. Once the final version of the PTEPA rubric was developed, we re-rated the programs and shared the results with program leaders for a second review. The PTEPA rubric was modified through 22 versions to arrive at the current public version (version 2.0) [30].

C. Instrument review

In addition to the data from thriving programs, we triangulated the PTEPA rubric with other sources of evidence, including systematic review by three experts in PTE and comparison to other instruments and related reports and literature. As described later in Sec. V, these reviews provided evidence of the instrument's validity (both organization and content), as well as contributing to some additional modifications of items and structure.

D. Ethics

In developing the PTEPA rubric, we paid particular attention to methods intended to uphold ethical integrity. While ethical considerations are always paramount, the development of a program assessment rubric has particular potential to harm study participants and the community at large. For example, if PTEPA rubric ratings for a program were made public, this could harm the institution (by reflecting poorly on their activities) or a program leader (by damaging relationships if they are perceived as having rated their colleagues or collaborating units poorly). The instrument as a whole also has the potential to do harm if it holds up unrealistic expectations for the PTE program community, reflecting negatively on good efforts and damaging morale and support. Guided by the American Evaluation Association's "Guiding Principles for Evaluators," [37] we attempted to address these ethical issues and protect the welfare and dignity of all stakeholders by

- Abiding by human subjects' research protocols, including protection of confidentiality.
- Publicly reporting only strengths (not weaknesses) of specific thriving programs.
- Seeking to include the range of relevant perspectives of the stakeholders in the work by interviewing experts, researchers, students, alumni, program leaders, administrators, Teachers in Residence, and so on.
- Ensuring that findings and outputs were justified by the data, attending to the needs and welfare of the American Physical Society as the director of the work, as well as the broader PTE program community, and upholding principles of high-quality research.
- Attempting to support the most appropriate use and interpretation of the instrument, emphasizing that rubric results are not to be overinterpreted (e.g., one does not assign numerical "scores" to rubric results) and that the rubric is intended to be used only for self-assessment rather than for external review.
- Creating supports for using the rubric effectively and for data visualization of the results to maximize the benefit to the community.

IV. RUBRIC STRUCTURE

One major output from this study is the creation of the PTEPA rubric itself. In this section, we describe the results of the work in terms of the development of that final rubric. In the next section, we describe what a teacher education program team can learn about itself by engaging in self-study using the rubric.

A. Standards and components

The structure of the rubric underwent significant revision over time. Our main goal was to create an organization that was meaningful and useful, keeping the instrument to a manageable length, and each section of the rubric a somewhat similar length. The first draft used five main categories that were similar to the five categories of the TEPA instrument, except that "Beginning teacher support" and "Teacher and school development" on the TEPA were not included in the PTEPA rubric as they were not aligned with our focus; instead, "Professional community" and "Targets, tracking and program assessment" were added as distinct categories. Other categories on the original PTEPA rubric included "Infrastructure, policy, leadership, and collaboration," "Physics content and physics pedagogy," and "Recruitment and retention." A few of these categories were divided into subcategories. The final organization of the PTEPA rubric (see Table I) bears some similarity to

these original categories but reflects changes made as a result of instrument validation (see later in this paper).

The main areas of the PTEPA rubric that were significantly developed beyond what existed in the TEPA were leadership, the collaboration between units, and the role of the physics department in recruiting and preparing teachers. At the end of all thriving program visits, due to addition of items in these and other areas, we ended up with a total of 8 categories: infrastructure and policy, leadership and collaboration, the physics department, pedagogy and curriculum, recruitment and retention, mentoring, professional community, and targets, tracking, and program assessment. The "physics department" category captured valuable elements not originally on the TEPA (such as the number of majors, quality of the physics courses, strength of the student community, advising processes, and faculty encouraging students to pursue careers in teaching), based on other work this area [1,14,17,38]. However, we felt that a total of 8 categories would be overwhelming to a user, and some categories had only a few items in them. Thus, after all program visits had completed, we combined the categories mentoring and professional community, since they have similar goals (to support successful degree completion and retention in the profession). The physics department category was also problematic: it included items (such as an LA program and the number of physics majors) that had quite different ultimate purposes (such as recruitment, physics content knowledge, or mentoring), despite their common home within the physics department. Thus, we redistributed items in this category into other appropriate categories reflecting these purposes, resulting in the final set of 6 main categories. These six categories, eventually termed standards, reflect the narrative arc of a PTE program, beginning with a solid institutional foundation and proceeding to bring students into the program, prepare them, and assess programmatic success (see Table I).

In naming these categories (and subcategories) of the rubric, we felt that using the language of accreditation would be valuable, to indicate that these are suggested performance indicators (even though there is no intention to accredit programs). For teacher educators, the Council for the Accreditation of Educator Preparation (CAEP) evaluation rubric [39] is a well-known instrument for accreditation of broad teacher education programs, and we found its structure easy to understand. Thus, we borrowed the CAEP rubric's language of standards and components, as well as giving narrative descriptions of those standards and components.

B. Items

The first draft of the PTEPA rubric was heavily based on the TEPA instrument and had a total of 35 items. The current version of the PTEPA rubric has 89 items; the increased number of items reflect physics-specific areas that were not reflected on the TEPA and the wide variety of elements that were seen as important at the thriving programs. Additionally, many TEPA items also addressed multiple elements of teacher preparation within the same item (such as different types of professional communities, or a variety of data that might be tracked), and these elements were separated into individual items. The Supplemental Material [40] outlines all the items that were added to the PTEPA rubric that were not originally on the TEPA (such as Arts and Sciences-level support for teacher education, collaboration between physics and education on the licensure pathway, or the number of physics majors). This Supplemental Material catalogs the highly physics-specific elements of the PTEPA rubric [40].

In addition to creating new items, we often modified items that came from the TEPA: e.g., "promotion and tenure" became "promotion and tenure in physics," and physics-specific elements of clinical practice such as the disciplinary expertise of the university supervisor were emphasized. Scale points were added to all items and revised over time. Detailed histories of these and other items on the PTEPA rubric are available in the report, *A Study of Thriving Physics Teacher Education Programs* [36].

Several items that were originally on the TEPA, or were developed in the course of iterating the PTEPA rubric, were later removed. Typical reasons for removal were a lack of a theoretical or empirical basis for considering that item to have a positive influence on physics teacher production, inability to clearly define the item, or determining that the item lay outside the scope of the rubric. Examples of removed items are given in Table II and in Appendix 6 of Ref. [36].

Items on the PTEPA rubric are intended to be unidimensional and independent (as much as possible) with scale points that increase ordinally (though we do not claim linearity). For example, a single item about Learning Assistant (LA) programs was eventually separated into items measuring the availability, attractiveness, and participation in LA programs. This attention to unidimensionality also increased the overall number of rubric items.

C. Item scale points

We originally developed four levels of achievement (scale points), plus an "absent" level, but this was found to be overly complex; we found three clearly defined levels (with the middle level denoting a "sufficient" level), plus not present, to be more usable. This decision is in alignment with Bresciani, Zelna, and Anderson [41], who indicate that it is harder for a reader to make sense of more than three levels. Additionally, three levels are typical for evaluation rubrics [26]. The CAEP Rubric [39] uses levels that are "below," "meeting," and "above" a sufficient level, giving "examples" of attributes at each level. This approach informed our language for our scale points: benchmark is considered the recommended level, with developing below that level and exemplary above it. Also, our language

TABLE II. A sample of items removed from the PTEPA rubric.

PTEPA rubric item	Reason for removal
PTE program position	Described whether there was a named position (such as "program coordinator"), as advocated in Ref. [1]. Removed because some program leaders held such a title and others did not, and it was unclear that there was any effect of this designation.
Support for physics teaching improvements	Described the climate in the physics department for faculty making changes to their teaching, which is an element of a strong physics program [14,15]. Removed because the quality of the introductory course has a more direct influence on potential teacher candidates.
Monitoring for student success	Originally on the TEPA instrument; measured whether the program monitored student progress. Removed because the college or school of education typically includes systems for tracking student progress toward requirements, and so this item was not specific to PTE programs. However, items describing PTE-specific mentoring were maintained.

reflects our belief that even programs that are just beginning to support PTE are doing more than a typical U.S. institution and deserve positive feedback and support (e.g., developing vs "below sufficient level"). The definitions of our scale point descriptions reflect a philosophical stance of recognizing all achievement as important while pushing programs towards excellence. The scale point descriptions are as follows:

- *Not present*: The item is not present in the program or does not meet a minimum level.
- *Developing:* The program performs better than a typical U.S. institution of higher education on that item.
- *Benchmark:* The program performs at a recommended level on that item.
- *Exemplary:* The program is among the best performing on that item.

D. Item prevalence

With this abundance of items, we were concerned that users would have difficulty prioritizing activities or making sense of results. We were unwilling to drop items from the instrument for simplicity, however, as it is yet unclear how broadly important each might be, and for which contexts. We acknowledge Lorrie Shepherd for providing us with the idea of distinguishing items that are more (or less) critical. Using results across all 8 programs, we identified the most commonly strong items as "essential" and the others as "enabling," intending to reflect whether items were central to success versus simply supportive. However, we determined that this wording implied a causality that could not be supported by the data available. Items may be commonly observed because they are easy to achieve, because they are precursors of success, or because they are required for success. Thus, we changed the wording to identify "prevalent" items, reflecting simply the empirical observation that they are more commonly observed. Items that are not prevalent receive no label. Our criteria for "prevalence" was that at least 75% (six) of the studied programs were rated at least benchmark (the recommended level) on that item. Exceptions were made for a few items about an undergraduate teacher certification route, as some institutions included only a post-baccalaureate program. For those items, five out of the six programs with undergraduate certification pathways were required to meet the benchmark level.

Because there are only 8 studied programs, a rating that is inaccurately high at even a single program could result in an item inappropriately labeled as prevalent. Thus, each item was required to pass one of the following "confidence" criteria to be considered as prevalent:

1. The item is inherently reliable because it measures an objective quantity, such as the number of faculty leaders or Teachers in Residence, or
2. At least six of the eight studied programs were rated exemplary on that item, or
3. Both members of two pairs of very different types of programs were rated at least benchmark on that item, indicating that the item is important across contexts (e.g., the item was rated at least benchmark at two types of programs, such as a large STEM teacher education program run from outside the physics department *and* a small program led by a single physics faculty member).

This process resulted in about 50% of PTEPA rubric items being labeled prevalent: a total of 44 items, out of 89 on the instrument as a whole. The standard with the greatest proportion of items labeled as prevalent is Standard 2: Leadership and collaboration, in which 16 items (or 76% of the items in the standard) are labeled prevalent. Other standards typically have about 30%–45% of their items labeled as prevalent. As data are gathered from additional programs, we will be able to determine the degree to which these items are broadly prevalent in strong programs, and the degree to which they do (or do not) predict the number of future physics teachers produced by programs.

V. VALIDITY AND RELIABILITY

We have investigated several forms of the validity and reliability of the PTEPA rubric. Rubric data are only available for eight "thriving" physics teacher preparation

TABLE III. Sample validation of rubric components and items.

PTEPA rubric element	Content validity argument
Component 2C: Program collaboration	The importance of collaboration between academic units is documented in several reports [1,42]. The specific items in this component were largely developed anew for the PTEPA rubric to describe the possible intersection points between physics and education and characterize the health of that intersection. While there is some overlap among elements included here versus other areas of the rubric (such as student teaching or advising), the items in this component explicitly describe the *collaboration* on this element as the dimension of interest. We explicitly define collaboration as between the program and relevant units (wherever the program is housed) to allow for the diversity of programs observed (i.e., some are run through the physics department, some through the college or school of education, and some through an independent unit).
Item 2B-3: Disciplinary expertise	This item captures the importance of having both disciplinary and pedagogical expertise on the team. While reports [1,15] refer to the importance of having collaborators in physics and education, we observed solo program leaders who had both types of knowledge (our benchmark level). We also noted that project collaboration and communication was enhanced in programs whose leadership represented more than one department, or whose leaders spanned multiple departments and that this effect was separate from their disciplinary expertise. Thus, we differentiated between "disciplinary expertise" (this item), "boundary crossers," and "departmental representation" (both in 2C: Program collaboration).

programs; additional data, especially from nonthriving programs, will enable us to address additional areas of validity. In this section, we discuss the degree to which the rubric has demonstrated different forms of validity and reliability for its intended purpose: *Does the PTEPA rubric measure features that thriving programs tend to have?*

A. Substantive validity

Is there a good theoretical basis for the features included on the PTEPA Rubric? Substantive validity has been demonstrated by the fact that the PTEPA rubric is based on the TEPA, which underwent significant development through literature review, program visits, and focused input from experts and practitioners. The extensive development and validation process of the TEPA is documented in Ref. [34]. Additionally, we conducted a careful cross reference of the PTEPA rubric items with notable reports in physics teacher education (see Appendix 5 of *A Study of Thriving Physics Teacher Education Programs* [36], finding that elements on the rubric were supported by at least one (and usually more) of the reports we had identified as relevant [1,14–17,27,35,39]. We note that the PTEPA rubric was not particularly designed to align with most of these reports: the fact that empirical results led to standards, components, and items that aligned well with what is already known provides some validation of the content and organization of the instrument.

B. Content validity: Items and components

Does the PTEPA rubric include all the features that are important for thriving programs? In addition to the substantive validity described above, the content of the rubric has been validated empirically through visits to thriving programs and expert review, as well as a literature review cross referenced with rubric items. These activities often resulted in the addition of items that were seen as missing (such as whether the program team has a coherent vision of teacher education), or removal of items whose effects could be either positive or negative (such as whether the program had autonomy and functioned as a separate unit). Examples of the demonstrated content validity for components and items are provided in Table III. See Appendix 6 of *A Study of Thriving Physics Teacher Education Programs* [36] for a similarly detailed history of rubric elements.

We consider the standards to be a helpful organization of items and components, but not necessarily a specifically valid element of the rubric. That is, we do not have a substantive basis for claiming that our standards organize the items and components into coherent constructs. The studied programs demonstrated variable performance on elements within the standards, and thus the standards may not arise empirically from the data (e.g., as in factor analysis). Components, however, are more tightly focused. In examining the data from the 8 thriving programs, we find that a majority of components (11 out of 19) show consistent results across a majority of items for a majority of programs. We define consistency in this case as having items ratings that are over 75% benchmark or exemplary (consistently strong), or over 75% developing or not present (consistently weak). For 11 components, at least five of the studied programs meet these consistency criteria. For the other 8 components, 3 or 4 programs showed mixed results in those components (usually around 50% items rated strong and 50% rated weak). These results suggest that the items in some components form a coherent construct, but additional data may allow stronger modeling and validation.

The fact that half of the items on the rubric were identified as prevalent also provides evidence that the rubric measures

what thriving programs do. However, it is still possible that important items are missing; collecting data from additional programs (including different types of programs, such as those at smaller institutions) will help expand the content validity of the rubric. One question we want to be able to eventually address is the degree to which certain elements of the rubric are critical, and whether this depends on the type or size of the institution or program.

C. Interrater reliability

Do independent raters agree on rubric ratings? We have not undertaken a detailed reliability study, and such a study is unlikely, given the amount of time it requires for any researcher to become familiar enough with a program to rate it. That said, we have confidence in the reliability of the ratings of the 8 thriving programs, as these ratings were discussed among researchers and with program leaders, resulting in consensus in most cases. Additionally, the scale points were revised many times in an explicit attempt to increase reliability by continually revising item wording, or distinctions between levels, to enable consistent interpretation among program leaders, experts, and researchers. For example, in some items, a program was originally able to achieve exemplary level without also achieving benchmark, or they might achieve some elements of the exemplary level but not all. In these cases, the item was reworded to help program leaders more clearly identify exactly which level they had achieved.

D. Self-rating reliability

Do program leaders self-ratings match with those of trained observers? Program leaders exhibit some variations in their rating tendencies. Some tend to be more conservative, only accepting ratings at a level supported by evidence, whereas others are more optimistic in their ratings, tending to portray their program in a positive light. A complicating factor is that, in most cases, a single person does not have all the information necessary to complete the rubric. To enhance reliability and provide more insight into areas of discrepancy, we developed a process for programs to use in completing the PTEPA rubric:

1. Several members of the team complete the rubric individually, taking notes to justify responses.
2. The team comes to consensus on ratings.
3. A trained researcher provides feedback based on the team's notes and a reflective interview.

This process has worked well for a few programs already. As possible, we are gathering data on the individual versus consensus responses to enable us to identify which areas (and how many) tend to be most unreliably rated by individuals. We will also do think-aloud interviews with some programs to better understand and mitigate reliability issues. Lastly, some programs will be asked to complete the rubric again in 6 months (during which their program would not be expected to undergo significant changes), to check for test-retest reliability.

As additional data are gathered from a wide range of programs, we hope to be able to address the degree to which the PTEPA rubric demonstrates the following types of validity:

- Criterion validity. *Does a high score on the rubric relate to high levels of physics teacher production?* We will collect data from thriving and nonthriving programs, and determine if rubric results predict levels of teacher production (likely using multiple regression, nonparametric modeling study that controls for the size of student population).
- Predictive validity. *Does increasing the rubric score lead to increased numbers of physics teachers produced?* We are asking programs that are newly engaged in PhysTEC to complete the rubric at the start of their engagement. Comparing these results to rubric results over time will help identify to what degree there is a correlation between *increases* in rubric ratings and subsequent *increases* in teacher production.
- Process validity. *To what extent do program leaders understand the rubric items as they were intended?* The reliability studies outlined above will help to address this question.
- Consequential validity. *To what extent are the rubric results used in ways that improve local programs? The* reliability studies, as well as data gathered on how the rubric is used, will help to address this question.

VI. HOW TO USE THE RUBRIC

The PhysTEC project suggests that programs use the rubric collaboratively with others on their program team, or engage other stakeholders as needed. Completing the rubric can be a good way to start a conversation with the college or school of education about certification options, for example. Some faculty have also planned to use it to advocate for resources internally or externally, by identifying strengths, gaps, and improvements in the program. We suggest completing the rubric on an annual basis: this is supported in the interactive Excel rubric through an "annual review" tab, where results from a previous year can be compared to the current year. The intention is for the rubric and its associated supports to assist a PTE program team in engaging in a process of continuous improvement, identifying areas for growth and reflecting on their progress. Feedback has already been positive in this regard, with program leaders in physics or education using the rubric as an opportunity to engage with partners in other units, and to guide meetings with stakeholders, including deans and other administrators. Possible opportunities for completing the rubric include the following:

- When applying for funding (e.g., Noyce, PhysTEC).
- When preparing to make a case to administrators for PTE program funding.

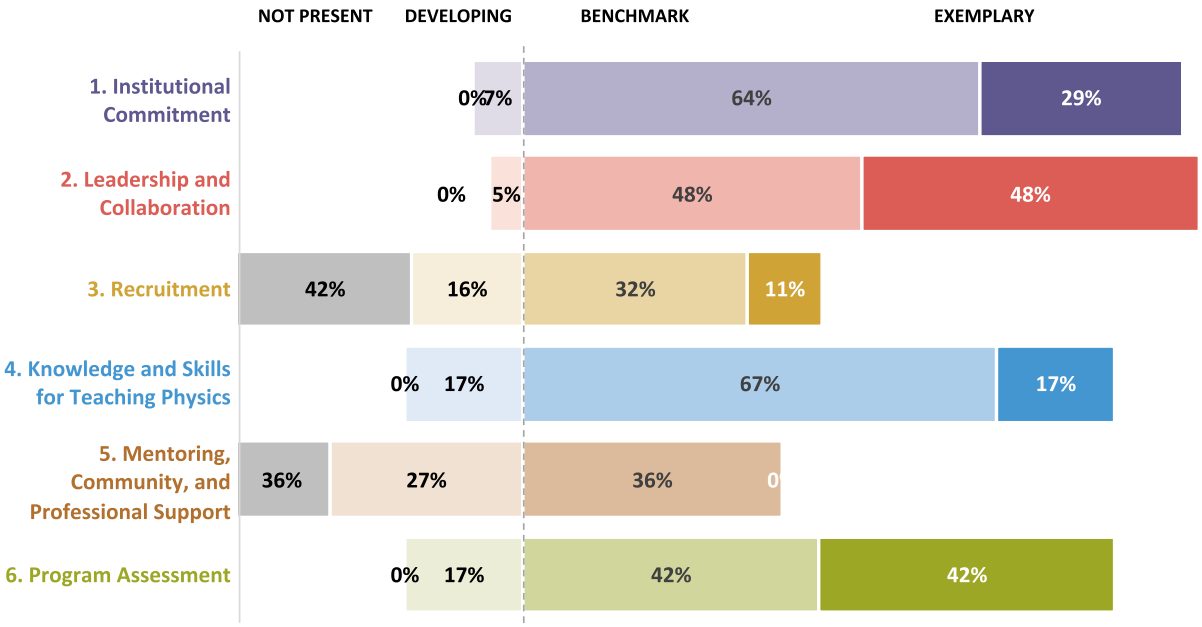

FIG. 1. Sample program data for the PTEPA rubric across the 6 standards, showing the percent of items rated not present (gray), developing (light shade), benchmark (medium shade), or exemplary (dark shade). Note that different standards have different numbers of items. Percents may not add to exactly 100% due to rounding.

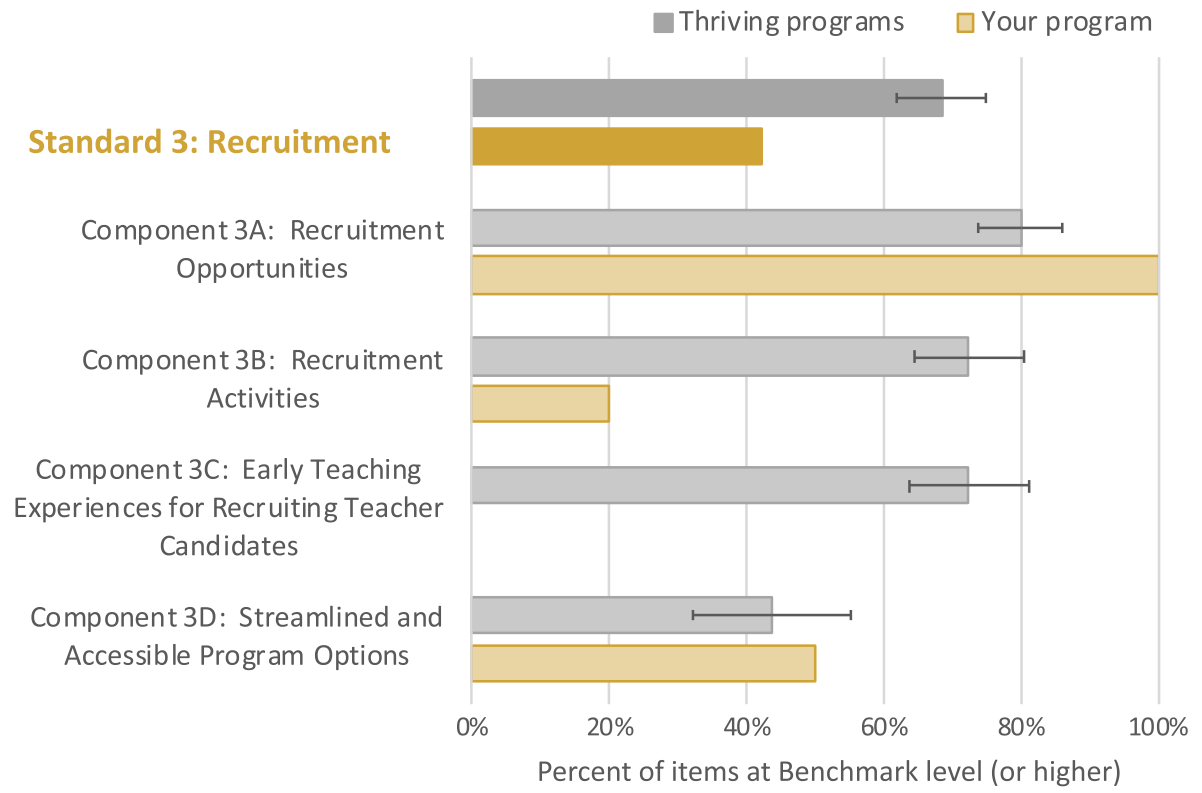

FIG. 2. Sample institutional data for the PTEPA rubric for standard 3 for a fictitious program, compared to the actual data for the 8 thriving programs. Percentages represent the percent of items rated at benchmark or exemplary level. Thriving program data represent the percentage at least benchmark, averaged across the items in that component and across the 8 thriving programs. Error bars represent the standard error of the mean.

- When preparing annual reports on the PTE program.
- During PTE program planning or review.
- During department strategic planning or retreats.

A. Interpreting results

The ratings do not have numerical values, and item ratings should not be added across the rubric to calculate a score. Such numerical scoring would not be valid because rubric items and components are neither independent nor of equal weight. For this reason, all visualizations and synthesis of rubric ratings that we provide focus on the ratings themselves (e.g., the level at which an item was rated, or the percent of items rated at different levels across a standard or component). Figure 1 (previous page) shows an example of a visualization of data from a fictitious PTE program. The percent of items rated at least benchmark level appear on the right side of the dotted midpoint line, and those that are not yet at benchmark level appear on the left side of the line, allowing a quick visual assessment of strength and weakness.

Program leaders should not try to maximize ratings in all areas of the rubric, as this is not realistic. Even thriving programs are not uniformly strong in all areas of the rubric [28,33,36]. Additionally, not every element of the rubric is important for every institution, or at every point in time in the lifetime of the program.

That said, this PTE program might find it valuable to look at standards 3 and 5 in detail to learn if there are weaknesses worth addressing. Figure 2 (previous page) gives the percentage of items that meet at least the benchmark level of achievement for standard 3, compared to the 8 thriving programs in the study. Fewer items are rated at the benchmark level in components 3B, 3C, and 3D for their program than for the thriving programs. While the comparisons to thriving programs are an imperfect comparison, they still provide a useful (and the only) baseline for comparison.

We find it valuable to combine an investigation of strength and weakness at the component level with an investigation of particular items leading to that overall strength and weakness. Figure 3 (following page) shows the report that program leaders are given in the Excel version of the PTEPA rubric. The spreadsheet also includes a user-fillable section to indicate an "action plan" for each component.

The PTEPA rubric User's Guide also provides a worksheet for "importance and synthesis ratings," where users can rate the importance (low, medium, high) of each standard and component for their program (1) now and (2) in three years. Users may also give a synthesis rating, indicating whether they consider their overall achievement on each standard and component to be poor, fair, good, very good, or excellent, given their institutional context. Figure 4 (following page) gives an example of what this report might look like for standard 3 for this program.

If the program leaders make changes to the program and complete the rubric again in a year, the annual review portion of the interactive Excel rubric visualizes changes to the results, as in Figure 5.

B. Getting support

As described in Sec. III D, we consider it our duty as researchers to support effective use of the rubric for the good of the PTE program community and society overall. We have invested considerable effort in making the rubric usable and supporting that use through a variety of tools [30], including

- "Getting started with the PTEPA rubric," a 2-page handout with 10 broad questions about physics teacher preparation at an institution, to be given to a dean or potential collaborator to generate interest in improvement.
- A "PTEPA party" handout, to be given to stakeholders to advertise a gathering to collaboratively complete the rubric and discuss results.
- A fillable PDF version of the rubric.
- An interactive Excel version of the rubric, which automatically generates data visualizations.
- Additional visualizations made available when users share their data with PhysTEC evaluators.
- A "snapshot" version of the rubric without scale points to enable a glance at the structure.
- The "PTEPA rubric User's Guide," including questions that can be used to gather information on the program, a worksheet for rating the importance and strength of different rubric elements, and a narrative self-study template to document results and action plans.
- Workshops, webinars, online office hours, and individual meetings to support institutions in completing the rubric and interpreting results.
- Broad dissemination of the report and rubric in a variety of channels.

VII. DISCUSSION AND LIMITATIONS

A. Summary

To support the development of high-quality PTE programs and thus the education of more highly qualified physics teachers, we developed a rubric that provides an objective measure of features that are commonly observed at thriving PTE programs (those that consistently graduate 5 or more highly qualified physics teachers each year). The resulting analytic rubric consists of 89 items, organized into 6 standards, each with three levels that are described in detail for each item: developing, benchmark, and exemplary (plus not present). Items that are consistently strong at the studied programs are identified as prevalent. The study abided by a variety of ethical principles to avoid negative repercussions of rating PTE programs.

The rubric items and their levels were based on existing instruments, plus program visits to 8 diverse thriving

Standard 3: Recruitment	42%	
Component 3A: Recruitment Opportunities	**100%**	
3A-1 Physics majors	EXEMPLARY	★★★
3A-2 Physics-aligned majors	BENCHMARK	★★☆
3A-3 Physics teaching advisor	BENCHMARK	★★☆
3A-4 Recruitment network	EXEMPLARY	★★★
3A-5 Program identity and reputation	BENCHMARK	★★☆
Component 3B: Recruitment Activities	**20%**	
3B-1 Physics teaching ambassador	DEVELOPING	★☆☆
3B-2 Accurate information about career benefits of teaching	NP	☆☆☆
3B-3 Program promotion	BENCHMARK	★★☆
3B-4 Physics faculty discuss teaching as a career option	DEVELOPING	★☆☆
3B-5 Physics department exposes students to diverse career options	NP	☆☆☆
Component 3C: Early Teaching Experiences for Recruitment	**0%**	
3C-1 Attractiveness of early teaching experiences	NP	☆☆☆
3C-2 Exposure to intellectual challenge of teaching	NP	☆☆☆
3C-3 Availability of early teaching experiences	NP	☆☆☆
3C-4 Recruitment within early teaching experiences	NP	☆☆☆
3C-5 Exposure to K–12 teaching environments	NP	☆☆☆
Component 3D: Streamlined and Accessible Program Options	**50%**	
3D-1 Undergraduate licensure pathway	BENCHMARK	★★☆
3D-2 Post-baccalaureate licensure pathway	BENCHMARK	★★☆
3D-3 Time to certification for physics teacher candidates	DEVELOPING	★☆☆
3D-4 Financial support for physics teacher candidates	NP	☆☆☆

FIG. 3. Item-level report from the PTEPA interactive rubric (Excel format). Shaded items are prevalent (see Sec. IV D). Percentages shown are the percent of items within that component or standard that are at least at benchmark level.

programs. The rubric items were developed to focus specifically on the areas that are specific to physics teacher education (rather than teacher education in general). During the study, items were added or removed, and levels revised, to reflect this focus, and the features observed at the 8 thriving programs. Over time the items were organized into standards and components; this language reflects accreditation terminology, indicating that the rubric provides recommended performance indicators. Additionally, items that were observed to be

Standard or Component		Importance *Lower – Medium – High*		Synthesis Rating *Poor – Fair – Good – Very Good – Excellent*		
		Now	In 3 years	Year 1	Year 2	Year 3
3	RECRUITMENT					
3A	Recruitment Opportunities	High	High	Excellent		
3B	Recruitment Activities	Medium	High	Poor		
3C	Early Teaching Experiences for Recruiting Teacher Candidates	Low	Low	Poor		
3D	Streamlined and Accessible Program Options	High	High	Fair		

FIG. 4. Example importance and synthesis ratings table, from PTEPA User's Guide.

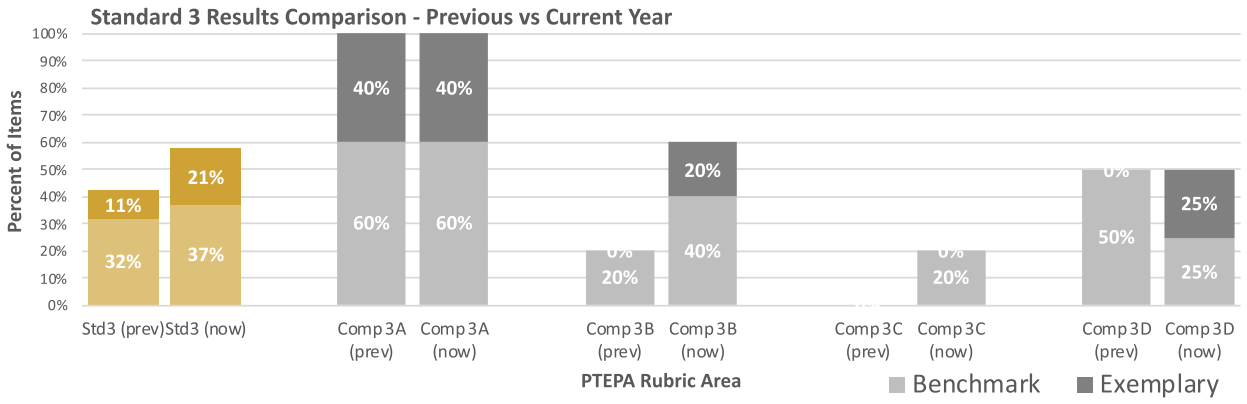

FIG. 5. Comparing PTEPA rubric results for the previous year to the current year for standard 3. The percent of items at the benchmark and exemplary level are shown.

at least benchmark level at 75% of studied programs, and that met one of three confidence criteria, were denoted as prevalent.

The rubric has demonstrated a variety of forms of validity for its intended purpose; that is, to measure features that thriving programs seem to have. It is based on a valid instrument (the TEPA), the features of the PTEPA rubric are well aligned with existing reports in PTE, and items were validated empirically through visits to thriving programs. The studied programs demonstrated diversity in their ratings on the items, standards, and components of the rubric (with a few exceptions).

PhysTEC recommends that the rubric be completed as a collaborative exercise: possibly as an opportunity for regular self-reflection, to engage with partners in the college or school of education, to guide meetings with administrators, or to create a case for program resources. Many resources and data visualizations support the use of the rubric for these and other purposes.

B. Limitations

Only eight programs were studied for rubric development: all are programs at large institutions, and most are mature programs with an undergraduate pathway (only two programs relied only on a post-baccalaureate or master's program for licensure). All eight programs are U.S. institutions; this aligns with the purpose of the study, but limits our findings to PTE in the U.S. The rubric is less suited to smaller institutions, or those without an undergraduate pathway. Programs at such institutions should especially beware of comparing themselves to thriving programs, and should instead use the PTEPA rubric as a guide of areas to be considered to strengthen a program.

Another limitation is in the basis for the organization of the rubric. Additional data are needed to learn whether standards and components in the rubric represent coherent and reliable constructs.

Because of the complexity of the rubric, researchers are likely to need to rely on program leader self-ratings on the rubric. This presents another limitation, as the authors have already observed variability in self-ratings. We attempt to mitigate this limitation by requiring consensus ratings.

Additionally, the length of the rubric has deterred some leaders from using it. More data from diverse programs are required to be able to further develop the rubric to better capture the range of programmatic practices *and* to attempt to reduce the length of the rubric to focus on the most critical items.

C. Next steps

We foresee four main next steps for the rubric:

1. Encourage and support rubric use,
2. Conduct further validation studies,
3. Further develop the rubric, and
4. Use the rubric in research studies.

1. Encourage and support use

We have developed several supports for rubric use, and PhysTEC is investing resources in advertising the rubric (through mailings of the physical report, newsletters, emails, a webpage, webinars, and so on.) PhysTEC-supported programs are now required to complete the rubric; additionally, their action plans are required to address gaps in the rubric, and the annual reporting structure will ask them to reflect upon their progress in rubric terms. For all PTE programs (either supported by PhysTEC or not), it will be valuable for the community to develop even more knowledge and information-sharing opportunities so that program leaders can develop the most effective solutions to gaps identified in their programs using the rubric.

However, even faculty who are motivated to improve PTE may not prioritize self-assessment if they are not held accountable for doing so by PhysTEC funding. One implicit incentive is that the rubric results can be used to argue for resources internally (e.g., the dean) or externally (e.g., in Noyce grant applications). More explicit incentives

TABLE IV. Planned validation activities.

Type of validity	Associated evaluation question	Potential or planned activity
Content validity	How does the rubric need to be modified for smaller institutions? What is missing?	Continue to administer the rubric at a variety of institutions and modify over time.
Reliability	Does the PTEPA rubric work for self-study as well as when used by trained observers?	Conduct 6-month test-retest reliability with a subset of respondents. Compare consensus PTEPA rubric responses to individual responses.
Process validity	To what extent do program leaders understand the PTEPA rubric items as they are intended?	Interview program leaders as respondents on their interpretation or understanding of rubric items.
Consequential validity	To what extent are PTEPA rubric results used in ways that improve local programs?	Interview program leaders on their programs' response to rubric evaluation.
Criterion validity (concurrent and predictive)	Do different PTEPA rubric results (1) correlate with and (2) predict different rates of physics teacher production? Which items on the rubric are particularly essential? In which context?	Apply the PTEPA rubric to thriving and nonthriving programs and conduct a multiple regression modeling study to investigate the relationship between PTEPA rubric scores and teacher production. Apply the rubric to a variety of PhysTEC-funded programs over time, including retrospectively, and analyze for a correlation with physics teacher graduation numbers.

might include a "PTEPA prize" for improvement on the rubric. After further validation of the rubric, it could be included as one of several pieces of evidence (such as program visit reports, student learning assessments, and program narratives) in identifying a standard of "excellence" akin to accreditation [43–45]

2. Conduct further validation studies

While the PTEPA rubric has demonstrated substantive and content validity, demonstrating both good theoretical basis for the items and alignment with empirical evidence, we plan to conduct further validity studies to better demonstrate its potential for use to answer important questions for the field (Table IV, above):

Some of these validity studies are within the scope of the current efforts, but others would benefit from broader engagement with researchers in PTE, especially for the collection of data from diverse institutions.

3. Further develop the rubric

As additional PTEPA rubric data are collected from programs, we are continuing to gather information about how users interpret the items and items that may be missing or misleading. These will help to inform additional revisions to the rubric over time. To help reduce overall rubric length, the validation studies noted above will help us to determine which items are more or less critical, and how these might depend on context, allowing us to make more informed decisions when considering items for removal.

4. Use it in research studies

Results from the rubric have significant potential to contribute to our understanding of PTE programs and other educational change initiatives. A companion paper [28] describes initial findings from the rubric, such as the observation that thriving PTE programs are strong in multiple areas of the rubric. Future research could help identify common program features in the U.S., the influence of local context on the features of strong programs, and the relationship between such features and PTE program graduation rates. Application of the rubric to non-U.S. programs can also help to elucidate the differences between the U.S. and other countries in terms of PTE. We encourage the PTE program community to submit their PTEPA rubric results to the authors of this study to enable us to conduct such studies, and we encourage the physics education research community to collaborate with us in this effort.

VIII. CONCLUSION

Through an extensive development process, we have developed a tool that

1. Characterizes the practices and structures observed at thriving PTE programs,

2. Provides a specific, objective, and reliable self-assessment for PTE program leaders, and
3. Supports research on PTE programs.

All these elements address the need laid out at the start of this paper, to support a goal-driven process of continuous assessment and improvement for PTE programs in the U.S. The rubric outlines the space of what is involved in effective PTE program design objectively for the first time. The rubric is not a universal checklist of standards for all programs to follow, but rather a guide of areas to be considered to strengthen a program. This tool has the potential to increase the number of high-quality physics teachers prepared annually in the U.S. by allowing program leaders to analyze their local landscape, choose strategies to address the local challenges, assess their results, and use results to guide future improvements, supporting critical thinking and local improvements to meet the national need in physics teacher preparation.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grants No. PHY-0808790 and No. PHY-1707990 and the APS 21st Century Campaign. We thank Monica Plisch, Director of Education and Diversity at the American Physical Society, for her leadership and intellectual contributions to the development of the rubric and interpretation of its findings. We thank the gracious hosts and teams at the programs visited for the study (described in Appendix A3) for hosting visits, participating in interviews, and reviewing the PTEPA rubric results and reports: Michael Marder (University of Texas at Austin), Valerie Otero (University of Colorado Boulder), Duane Merrell (Brigham Young University), Chance Hoellwarth (California Polytechnic University, San Luis Obispo), Brian Thoms (Georgia State University), Karen Magee-Sauer (Rowan University), Eugenia Etkina (Rutgers University), and Keith Sheppard (Stony Brook University). Additionally, we acknowledge the contributions of the following: Gay Stewart (West Virginia University), Stamatis Vokos (California Polytechnic University, San Luis Obispo), and Wendy Adams (Colorado School of Mines) provided an expert review and feedback on the PTEPA rubric from a physics teacher education perspective. Judy Oakden (Pragmatica) provided a review of the instrument from an evaluative rubric perspective. Claudia Fracchiolla (National University of Ireland Galway) created the interactive Excel version of the PTEPA Rubric. Justyna P. Zwolak (Joint Center for Quantum Information and Computer Science) provided data analysis and visualization, and Jessica Alzen (University of Colorado Boulder) provided feedback on data interpretation and conducted the SDI analysis. Anthony Ribera (Rose-Hulman Institute of Technology) provided a review of academic accreditation processes.

[1] D. E. Meltzer, M. Plisch, and S. Vokos, *Transforming the Preparation of Physics Teachers: A Call to Action. A Report by the Task Force on Teacher Education in Physics (T-TEP)* (American Physical Society, College Park, MD, 2012).

[2] I. V. S. Mullis, M. O. Martin, A. E. Beaton, E. J. Gonzalez, D. L. Kellly, and T. A. Smith, *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS) Center for the Study of Testing, Evaluation, and Educational Policy* (Boston College, Chestnut Hill, MA, 1998). Available at http://timss.bc.edu/timss1995i/HiLightC.html.

[3] M. Plisch, The Physics Teacher Education Coalition, in *Recruiting and Educating Future Physics Teachers: Case Studies and Effective Practices*, edited by C. Sandifer and E. Brewe (American Physical Society, College Park, MD, 2015), pp. 21–25.

[4] L. C. McDermott, A perspective on teacher preparation in physics and other sciences: The need for special science courses for teachers, Am. J. Phys. **58**, 734 (1990).

[5] E. Etkina, Pedagogical content knowledge and preparation of high school physics teachers, Phys. Rev. ST Phys. Educ. Res. **6**, 020110 (2010).

[6] E. Etkina, B. Gregorcic, and S. Vokos, Organizing physics teacher professional education around productive habit development: A way to meet reform challenges, Phys. Rev. Phys. Educ. Res. **13**, 010107 (2017).

[7] National Research Council, *Preparing Teachers: Building Evidence for Sound Policy. Committee on the Study of Teacher Preparation Programs in the United States, Center for Education. Division of Behavioral and Social Sciences and Education* (The National Academies Press, Washington, DC, 2010).

[8] National Science Teachers Association, Knowledge Base for Supporting the 2012 Standards for Science Teacher Preparation (2012). Accessed at http://www.nsta.org/preservice/docs/KnowledgeBaseSupporting2012Standards.pdf.

[9] J. G. Hill and K. J. Gruber, Education and Certification Qualifications of Departmentalized Public High School-Level Teachers of Core Subjects: Evidence from the 2007-08 Schools and Staffing Survey, *Statistical Analysis Report [NCES 2011-317]* (National Center For Education Statistics, U.S. Department of Education, Washington, DC, 2011). Available at http://nces.ed.gov/pubs2011/2011317.pdf.

[10] L. C. McDermott, P. S. Schaffer, and P. R. L. Heron, Preparing teachers to teach physics and physical science

effectively through a process of inquiry, in *Recruiting and Educating Future Physics Teachers: Case Studies and Effective Practices*, edited by C. Sandifer and E. Brewe (American Physical Society, College Park, MD), pp. 165–178.
[11] J. A. Marshall and James T. Dorward, Inquiry experiences as a lecture supplement for preservice elementary teachers and general education students, Am. J. Phys. **68**, S27 (2000).
[12] See http://phystec.org.
[13] PhysTEC Outcomes, accessed at https://www.phystec.org/Outcomes/.
[14] *Strategic Programs for Innovations in Undergraduate Physics: Project Report*, edited by R. C. Hilborn, K. S. Krane, and R. H. Howes (American Association of Physics Teachers, College Park, MD, 2003).
[15] R. Czujko, K. Redmond, T. Sauncy, and T. Olsen, *Career Pathways: Equipping Physics Majors for the STEM Workforce* (American Institute of Physics, College Park, MD, 2014).
[16] P. Heron and L. McNeil (Co-chairs), *Phys21: Preparing Physics Students for 21st Century Careers. A Report by the Joint Task Force on Undergraduate Physics Programs* (American Physical Society, College Park, MD, 2016).
[17] K. M. Aguirre, T. C. Balser, K. E. Marley, K. G. Miller, M. P. Osgood, P. A. Pape-Lindstrom, and S. L. Romano, PULSE Vision and Change rubrics, CBE-Life Sci. Educ. 12, 579 (2013). PULSE rubrics retrieved from https://pulse-community.org.
[18] S. Elrod and A. Kezar, *Increasing Student Success in STEM: A Guide to Systemic Institutional Change* (Association of American Colleges and Universities, Washington, DC, 2016).
[19] H. Altrichter, A. Feldman, P. Posch, and B. Somekh, *Teachers Investigate their Work: An Introduction to Action Research across the Professions* (Routledge, New York, NY, 2013).
[20] See http://ep3guide.org.
[21] *Recruiting and Educating Future Physics Teachers: Case Studies and Effective Practices*, edited by C. Sandifer and E. Brewe (American Physical Society, College Park, MD, 2015), pp. 165–178.
[22] *Teacher Education in Physics: Research, Curriculum, and Practice*, edited by D. E. Meltzer and P. S. Shaffer (American Physical Society, College Park, MD, 2011).
[23] Attributed to Peter Drucker.
[24] D. Allen and K. Tanner, Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners, CBE Life Sci. Educ. **5**, 197 (2006).
[25] J. King, K. McKegg, J. Oakden, and N. Wehipeihana, Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation, J. Multidisc. Eval. **9**, 21 (2013).
[26] E. J. Davidson, *Evaluation Methodology Basics* (Sage, Thousand Oaks, CA, 2005).
[27] C. R. Coble, L. DeStefano, N. Shapiro, J. Frank, and M. Allen, Teacher Education Program Assessment (TEPA): Assessing innovation and quality design in teacher preparation. (2012). Retrieved from http://www.aplu.org/projects-and-initiatives/stem-education/SMTI_Library/TEPA/file.
[28] R. E. Scherr, S. V. Chasteen, and M. Plisch, following paper, What do thriving physics teacher education programs do? Initial findings of the Physics Teacher Education Program Analysis (PTEPA) Rubric, Phys. Rev. Phys. Educ. Res. **16**, 010116 (2020).
[29] B. Pyper, analysis of 2018 Title II data (private communication).
[30] All PTEPA rubric materials are available at http://phystec.org/thriving.
[31] D. E. Meltzer, Research on the education of physics teachers, in *Teacher Education in Physics: Research, Curriculum, and Practice*, edited by D. E. Meltzer and P. S. Shaffer (American Physical Society, College Park, MD, 2011), pp. 3–14.
[32] S. K. Abell, International perspectives on science teacher education: An introduction, in *Science Teacher Education: An International Perspective*, edited by S. E. Abell (Kluwer Academic Publishers, New York, 2002), pp. 3–9.
[33] R. E. Scherr and S. V. Chasteen, Development and validation of the Physics Teacher Education Program Analysis (PTEPA) Rubric, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC* (AIP, New York, 2018).
[34] C. Coble, Developing the analytic framework: Assessing innovation and quality design in science and mathematics teacher preparation. White paper, Association of Public and Land-Grant Universities. (2012). Retrieved from http://www.aplu.org/projects-and-initiatives/stem-education/SMTI_Library/developing-the-analytic-framework-a-tool-for-supporting-innovation-and-quality-design-in-the-preparation-and-development-of-science-and-mathematics-teachers.
[35] PhysTEC Key Components. Retrieved from: https://www.phystec.org/keycomponents/.
[36] S. V. Chasteen, R. E. Scherr, and M. Plisch, *A Study of Thriving Physics Teacher Education Programs: Development of the Physics Teacher Education Program Analysis (PTEPA) Rubric* (American Physical Society, College Park, MD, 2018).
[37] American Evaluation Association Guiding Principles for Evaluators, Retrieved from https://www.eval.org/p/cm/ld/fid=51.
[38] M. Marder, C. M. Brown, and M. Plisch, *Recruiting Teachers in High-Needs STEM Fields: A Survey of Current Majors and Recent STEM Graduates* (American Physical Society, College Park, MD, 2017).
[39] *Council for the Accreditation of Educator Preparation*, CAEP accreditation handbook (Council for the Accreditation of Educator Preparation, Washington, DC, 2016). Retrieved from http://www.caepnet.org/accreditation/caep-accreditation/caep-accreditation-handbook.
[40] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.16.010115, which includes the version of the rubric referenced in this publication, interview protocols for development of the rubric, a list of items on the PTEPA rubric which do not appear on the TEPA rubric, a list of items removed from the PTEPA rubric, and ratings of all PTEPA rubric items for the 8 thriving programs in the study.

[41] M. J. Bresciani, J. A. Anderson, and C. L. Zelna, *Assessing Student Learning and Development: A Handbook for Practitioners* (NASPA—Student Affairs Administrators in Higher Education Washington, DC, 2004).

[42] S. Vokos and T. S. Hodapp, Characteristics of thriving physics teacher education programs, in *Recruiting and Educating Future Physics Teachers: Case Studies and Effective Practices*, edited by C. Sandifer and E. Brewe (American Physical Society, College Park, MD, 2015), pp. 3–19.

[43] National Institute for Learning Outcomes, accessed at https://www.learningoutcomesassessment.org.

[44] P. Pape-Lindstrom, T. Jack, K. Miller, K. Aguirre, J. Awong-Taylor, T. Balser, L. Brancaccio-Taras, K. Marley, M. Osgood, M. Peteroy-Kelly, and S. Romano, PULSE Pilot Certification Results, Letter to the editor, J. Microbio. Bio. Educ. 127 (2015).

[45] American Association for the Advancement of Science, SEA Change, Accessed at https://seachange.aaas.org.