A Guide to the Recommended Indicators and Measures for the Illinois Teacher Preparation Program Improvement and Accountability System

Prepared by

Teacher Preparation Analytics

In Cooperation with

Education First

Presented to the

Illinois State Board of Education

November 11, 2016

Table of Contents

About This Guide

This guide is intended to be used in conjunction with the Recommended Indicators and Measures for the Illinois Teacher Preparation Program Improvement and Accountability System, a document that reflects the work of the Illinois State Board of Education's (ISBE) Partnership for Educator Preparation (PEP) Steering Committee. The Steering Committee, a 20-member advisory committee composed of teacher educators, researchers, K-12 classroom teachers, state and district officials, and representatives of several important stakeholder groups, met from May-August of 2016 under the guidance of Education First (EF) and Teacher Preparation Analytics (TPA), who were retained as external facilitators and expert consultants to the project. The purpose of the Steering Committee's efforts was to recommend to ISBE a set of measures of the performance of the state's teacher preparation programs that reflect key outcomes in important program domains. This set of measures is intended principally to be used by ISBE and the programs themselves to identify program strengths and weaknesses and, to the extent possible, guide concrete steps toward program improvement. The measures are intended, as well, to provide meaningful and useful information to district officials, prospective teacher education candidates, and other stakeholders who have a keen interest in the quality and productivity of the state's educator preparation programs.

This guide provides several kinds of information that should aid understanding of the significance of each *indicator*[*] and the corresponding *measures*:

- More detailed descriptions of the indicators and measures
- Discussion of the importance of each indicator and measure
- Discussion of controversies or limitations surrounding the indicators measures
- Glossary of technical or unfamiliar terms

The Fundamental Question: Which Data Are Best to Use?

In pursuing the development of an annual teacher preparation program performance report as part of a larger program improvement and accountability system, Illinois joins 20 or more other states in the U.S. that have chosen to implement an annual teacher preparation program review intended to either supplement or supplant a more traditional, multi-year state accreditation or program approval process. (See the recent *CCSSO publication Accountability in Teacher Preparation: Policies and Data in the 50 States and DC*, which is available from the

[*] Words in blue italicized type are defined in the Glossary on the last pages of this Guide. The selected words will be italicized only one time in the text.

CCSSO website at http://www.ccsso.org/Documents/2016/50StateScan092216.pdf.) The stated objective of an annual review process varies somewhat from state to state, but its greatest promise is to promote a continuous improvement model in teacher preparation that involves a partnership in data collection, analysis, and strategic action between *education program providers (EPPs)* and the state agency or agencies responsible for ongoing approval of those programs.

There is a great deal of commonality in some elements of the data that are used by states engaged in annual teacher preparation program reviews. Some of these commonalities reflect data all states have been collecting for two decades in fulfillment of the requirements for Title II of the Higher Education Act. Very little of those data, however, are valuable for either preparation program improvement or accountability—although it is the express intent of the newly released Title II program reporting requirements that the new program reporting data be more suitable for these purposes. And as state agencies, the U.S. Department of Education, various advocates, and preparation programs themselves seek to respond to continued concern about the ability of programs to produce new teachers who overwhelmingly will be successful from the very beginning of their teaching careers, there have been new measures of preparation program performance that have been proposed and incorporated into some states' program review systems.

There is no body of scientific literature, however, that unequivocally supports the choice of specific measures that should be used for a valid and reliable review of educator preparation programs. The clearest articulation of this situation is a recent National Academy of Education report entitled *Evaluation of Teacher Preparation Programs: Purposes, Methods, and Policy Options*.[*] There is value in knowing the limitations of preparation program performance data, but many states and program providers nevertheless have felt a sense of urgency to move forward in developing new data systems for program improvement and accountability purposes.

In response to the absence of solid guidance available to states working to strengthen their program improvement and accountability systems, TPA in 2014 developed a set of Key Effectiveness Indicators (KEI).[†] Based upon the several objectives states were developing their systems to meet, TPA's challenge was to craft a limited set of essential indicators—and ultimately measures—that would (a) be useful for program improvement, (b) be equally applicable to all programs in a state (traditional and non-traditional) without prescribing

[*] M. J. Feuer, R.E. Floden, N. Chudowsky, and J. Ahn. (2013). *Evaluation of Teacher Preparation Programs: Purposes, Methods, and Policy Options*. Washington, DC: National Academy of Education.

[†] See M. Allen, C. Coble, and E. Crowe. (2014). *Building an Evidence-Based System for Teacher Preparation.* Washington, DC: Council for the Accreditation of Educator Preparation (CAEP). http://caepnet.org/knowledge-center

program content and structure, (c) be focused principally on program outcomes rather than inputs, (d) be compelling and transparent for a variety of stakeholders, and (e) respond to the concerns that both experts and the greater public have expressed about the effectiveness of programs and the caliber of individuals entering the teaching profession. Eventually we framed four categories of program assessment data and 12 indicators with accompanying suggested measures.

Insofar as possible, TPA has tried to ensure that the KEI are largely consistent with significant bodies of empirical research that have employed the indicators as reliable measures of teacher effectiveness (e.g., classroom observation) or predictors of teacher effectiveness (e.g., licensure examination scores). Other indicators, such as teacher placement and retention or program completion rates by sub-group, provide important information to various stakeholders about the extent to which individual preparation programs are serving the teacher needs of the individual states.

The KEI were the starting point for the work of the PEP Steering Committee, and they are reflected in the Recommended Indicators and Measures that emerged from the initial PEP process. The final decisions about the indicators and measures to be included in the Illinois program improvement and accountability process will emerge from further discussion and exploration as the work of the PEP now moves to a pilot phase. That phase may reveal that the data currently available in Illinois are inadequate for the immediate implementation of some of the indicators or measures. It may reveal that some of the indicators and measures are not indicative of important program strengths and weaknesses. And it may reveal that other indicators and measures are more appropriate to the task. The fundamental goal, in any case, is that the indicators and measures ultimately implemented effectively serve the dual need for continuous program improvement and fair and effective state program accountability.

Four Categories of Program Performance Data

The Recommended Indicators and Measures begin with the identification of four categories of program performance data: *Candidate Selection and Completion, Knowledge and Skills for Teaching, Performance as Classroom Teachers*, and *Contribution to State Needs*. These four categories correspond to those of the Key Effectiveness Indicators, and they were selected because they provide a multi-dimensional scan of teacher preparation programs that will be valuable both for indicating key factors that may be account for the strength of programs' performance on the various measures and for providing *triangulation* with other measures to achieve a more *reliable* analysis of programs' strengths and weaknesses. In addition, these four categories are applicable to any teacher preparation program, regardless of program structure,

so that ultimately the indicators and measures within the categories also will be applicable to any preparation program and thus provide a legitimate basis of comparison between programs.

Candidate Selection and Completion—This category may appear at first to be more a matter of program inputs than outputs. To be sure, the strengths, limitations, attitudes, experiences, and habits that characterize entrants into teacher preparation programs are not necessarily predictive of how well or poorly those entrants might teach once they enter the teaching profession. These data become important, however, as part of a larger longitudinal database that educator preparation providers (EPPs) and state agencies ought to maintain on each individual teacher candidate. Such a database will enable EPPs and individual *licensure* or *certification* programs to track the progress of each candidate through the program and in their professional careers beyond. Armed with such data, EPPs will be able to determine whether there is a significant correlation between specific strengths, limitations, and attitudinal characteristics of candidates and their success or difficulties in both the preparation program and their professional teaching careers. If any such correlations are discovered, this does not imply that programs should refrain from admitting candidates who have the limitations in question, but it may mean that programs do a disservice to those candidates (and to their potential P-12 students) unless they can provide greater support or remediation for them throughout their professional preparation.

The fact that teacher preparation programs do have various requirements for admission or require all applicants to go through a screening process, or that they have various admissions goals such as candidate diversity, also implies that the make-up of the teacher candidate pool for a particular program is a conscious outcome of those admissions criteria or goals. Differences in EPP practices concerning candidate admission should be respected insofar as possible, but not so far as to waive expectations for strong program performance.

Knowledge and Skills for Teaching—The data under this category provide the most direct evidence that the curriculum and enacted standards of a teacher preparation program ensure that candidates who complete the program possess knowledge of their teaching subject(s), an understanding of their role as teachers, and the teaching experience and skill required to be successful with their students as soon as they enter the profession. It is arguable that program completers' actual success as teachers is, at least to some extent, a function of factors in addition to their level of teaching skill and knowledge—the climate of the school in which they teach or the rate of absenteeism among their students, for example. It should not be arguable, however, that a preparation program's standards – as applied – either ensure or don't ensure that all candidates who complete the program (or *pathway*[*]) have the knowledge, experience,

[*] Some states allow pathways into teaching that don't require completion of a formal teacher licensure or certification program. These include, for example, state agencies' portfolio review of teacher candidates

and skill they need to be successful teachers. Since state standards for licensure and program completion are an important means to providing such an assurance to the public, these standards also come into play here.

Performance as Classroom Teachers—This data category responds to the driving question for an outcomes-oriented performance assessment of teacher preparation programs: How well do a program's completers perform as teachers in their own P-12 classrooms? In one sense, no other information about a program is more important than this because no matter how positive or negative a program looks based on all other data, if the program's graduates don't demonstrate skill and success in their real-world classrooms the program will be judged to be inadequate. As we noted above, however, several factors besides the program can influence a teacher's ability to be successful with her students. And, so, at least in the case of any individual teacher's classroom performance there remains the possibility that school or classroom characteristics or perhaps the placement of a teacher in a subject outside their teaching field thwart the application of the skills or knowledge the teacher may have gained in their preparation program.

In using teachers' classroom performance to assess the performance of their preparation programs, however, the data are not about any individual teacher, but about the multiple teachers who completed a particular preparation program. And the larger the number of teachers who contribute classroom performance data to the measures of program performance, the less likely the combined data will reflect anomalies and the more likely they will reflect the *mean* performance of teachers who completed a specific program. Over a large number of cases, the mean performance of completers becomes a reliable statistic, and if the mean differs significantly between two or more programs (especially when *statistical models* specifically account for school and classroom effects) this is prima facie evidence that programs with the higher scores are outperforming programs with the lower scores in important ways that should be investigated further.

One important caveat in relation to data in this category and the following category is that data for both categories will inevitably be incomplete because, at least at present, states and preparation programs are largely unable obtain the relevant information on program completers who teach in private schools or out of state.

Contribution to State Needs—The fourth program performance data category provides information of specific interest to teacher educators, university leaders, state and district officials, state legislators, and all others who have a stake in knowing what contribution individual teacher preparation programs make to the identified teacher needs of a state. In the

(sometimes as transfers from another state) or certification through an assessment by the American Board for the Certification of Teacher Excellence (ABCTE)

aggregate, the data from this category make it possible to construct a picture of how well a state's teacher production capacity aligns with its teacher needs. Using these data for a preparation program accountability system should not be taken to imply that all programs should aspire to address every kind of teacher need the state has. Indeed, private teacher preparation programs may not be under any mandate to respond to state needs, at all—although they are, by default, part of the available picture. And even public programs differ in their missions, in the interests and backgrounds of the student who enroll in their parent EPPs, in the regions of the state where their completers tend to take jobs, and in the percentage of out-of-state students they admit who are likely to return to their states of origin after they graduate.

Even allowing for these differences, the normative assumption—and the desired outcome from a teacher supply standpoint—is that most candidates who are prepared to teach in a particular state will enter the profession within a year or two of program completion in that state. And, given the documented ills of high teacher turnover,[*] it is also the desire on the part of state and district officials that candidates who enter the teaching profession stay in the profession for more than just a few years.

THE INDICATORS AND MEASURES

Under each of the four program performance data categories, the PEP Steering Committee has suggested a set of indicators—18 in all—for potential inclusion in the Illinois teacher preparation program improvement and accountability system. The indicators specify the various aspects of programs' performance that are to be measured. For most of the indicators, the Steering Committee has agreed to consider one or more actual measures of programs' performance. Some of these indicators and measures will be vetted in the upcoming pilot phases of the system development process, while others will be deferred for later consideration either because the data required to enact them will not be available anytime soon or because the consensus on the value of the indicators is less strong.

The companion document to this guide, Recommended Indicators and Measures for the Illinois Teacher Preparation Program Improvement and Accountability System, identifies the indicators and measures under consideration. It also lists the page numbers in the Guide that explain the rationale and some of the challenges for each of the designated indicators and measures.

The PEP Steering Committee used TPA's Key Effectiveness Indicators as a springboard for developing its own recommendations, and there are clear affinities between the KEI and the

[*] See, for example, R.M. Ingersoll (2001). Teacher Turnover and Teacher Shortages: An Organizational Analysis. *American Educational Research Journal* 38 (3), 499-534.

Steering Committee's recommendations. There are also important differences. Some of those differences are a function of the fact that the KEI combine indicators for the sake of economy that are more helpful for Illinois to consider separately. Some of them reflect priorities identified by the Steering Committee or ISBE. As for the measures, some of the differences are a response to the specific kinds of data available in Illinois or to the Illinois candidate and teacher assessment context. Regarding assessments of candidate knowledge and skill, for example, Illinois has identified target scores on licensure-related assessments that can replace *norm-referenced measures* proposed in the KEI with *criterion-referenced measures*. Illinois has also indicated a strong interest in increasing the number of minority teachers in its schools, and thus many measures include disaggregation by demographic categories such as race/ethnicity.

We now discuss the PEP Steering Committee's proposed indicators and measures in the order of their listing in the accompanying document.

Academic Strength—Although there is well-documented research showing that the academic proficiency of the U.S. teacher workforce has increased markedly over the past decade,[*] in a preparation program accountability context this indicator is highly charged. Efforts, such as those of the Council for the Accreditation of Educator Preparation (CAEP), the teaching profession's national accrediting body), to increase preparation program admissions standards have been met with strong protests from teacher educators and others that the consequence of such efforts will be to diminish the supply of teachers in the pipeline and exacerbate the difficulty of diversifying the teaching profession.[†] There is significant research, however, showing that teachers who are strong academically outperform teachers who are not.[‡] And there has long been concern—which undermines confidence in and respect for the teaching profession— that the teacher workforce in the U.S. has too many academically weak teachers in comparison with countries around the world recognized for their high-performing education systems.[§] Thus, both because of the importance of academic strength to teachers' classroom success and because of public concern about the adequacy of the teacher workforce, Academic Strength is an important program performance measure.

---

[*] See D.H. Gitomer. (2007). *Teacher Quality in a Changing Landscape: Improvements in the Teacher Pool*. Princeton, NJ: Educational Testing Service.

[†] C. Coble, E. Crowe, and M. Allen. (2016). *CAEP Standard 3.2 Research, Study and Analysis: A Report to the Council for the Accreditation of Educator Preparation*. (http://www.caepnet.org/standards/standard-3). In this CAEP-commissioned study on Accreditation Standard 3.2, TPA cites numerous instances of such protests and provides compelling evidence that raising academic standards for program entry could well diminish minority representation in the teacher workforce without vigorous efforts to recruit minority candidates who are academically more proficient.

[‡] Coble, Crowe, and Allen (2016).

[§] See B. Auguste, P. Kihn, and M. Miller. (2010). *Closing the Talent Gap: Attracting and Retaining Top-Third Graduates to Careers in Teaching.* New York City: McKinsey & Company.

Key issues in engaging this indicator are (1) how to measure academic strength and (2) whether the measures should apply to entering teacher candidate—possibly through criteria for program admission—or whether it's more important to gauge the academic strength of candidates upon program completion. CAEP Standard 3.2 suggests that academic strength measures can be applied either to entering or completing candidates. For entering candidates, the common measures are SAT or SAT scores for undergraduate programs, GRE or MAT scores for post-baccalaureate programs, basic skills test scores (Praxis Core, for example, claims to measure second-year college level proficiency), or Grade Point Average (GPA). The GPA measure might be high school GPA for program admission as freshmen, GPA in the first two years of college—including community college—for program admission as juniors, and the four-year college GPA for post-baccalaureate program candidates. Measures of academic strength for completing candidates would most likely include their college GPA, although scores on licensure assessments of candidates' content knowledge are also an indication of academic strength.[*]

There is research literature that supports or refutes the correlation of any of these individual measures with the eventual effectiveness of teachers in the classroom. Research also indicates, however, that the predictive *validity* of these measures is increased when they are combined, e.g., using both GPA and SAT/ACT.[†] From the standpoint of the IL program improvement and accountability system, the key question is whether and how these measures can be used as valuable data for program improvement purposes.

There are three important purposes of such measures:

1. To enable programs to determine how well academically stronger vs. weaker candidates perform in the program and, if possible, as teachers in the classroom—and based on that information to make decisions about the ability of the program to provide adequate support and remediation for candidates who are academically weaker upon program admission
2. To screen out candidates—either at admission or later in the program—who have academic deficiencies likely to be detrimental to their success as teachers
3. To ensure state officials and the public that program completers are academically sound.

For the purpose of tracking differences in the program and classroom performance between academically stronger and weaker candidates, it will be valuable to employ multiple measures at least in the early stages of implementation of the indicator in order to discover whether one

[*] Coble, Crowe, and Allen (2016).
[†] Ibid.

or more measures are particularly good predictors of program completion, performance in the programs, or performance in the classroom. This also will give programs a good indication of which thresholds for the measures are correlated with strong or weak candidate performance.

For the purposes of satisfying public concern about teacher quality, it is only those measures that permit a comparison between the performance of a program's teacher candidates and students outside of teacher education that will fill the bill. SAT, ACT and GRE scores serve this purpose most easily. In addition, for students seeking certification in a secondary education field, a comparison of their GPA in their major with the GPA of non-education students in the same major at the college or university can provide a reasonable teacher to non-teacher comparison.

The recommendation here is that the Illinois preparation program improvement and accountability system report out academic proficiency data on two *cohorts* each year. One cohort is the entering candidates in each program, and the other cohort is the exiting completers in each program. In addition, for program improvement purposes it may prove valuable to compare the academic proficiency measures for each candidate at program entry and completion as a means of ascertaining the relationship, if any, between the two measures.

Because both EPPs and individual certification programs have responsibility for ensuring the academic strength of their teacher candidates, performance measures under this indicator apply at both levels: at the individual program level (i.e. Elementary Education, Middle Grades Education, Secondary disciplines, etc.) and, in the aggregate, at the EPP unit level. District human resource directors and others who hire teachers, for example, may be specifically interested in the academic proficiency and teaching subject-related strength of candidates in specific certification fields. CAEP Standard 3.2 (the academic strength standard) applies to the EPP as a whole, and thus to reflect consistency with CAEP standards (a desire of ISBE) it is suggested that EPP-wide data on academic strength be reported.

Teaching Promise—This is largely an aspirational indicator because there are few reliable *standardized* assessments available that could provide a valid basis of comparison between candidates of different program. One such assessment that has recently been implemented in Missouri—though only to provide information on individual candidates' potential strengths and weaknesses and not for program accountability—is the Missouri Educator Profile, which was developed by NES Pearson.[*] Teach for America has designed its own, proprietary set of assessments of the teaching promise of potential program participants.[†] The Haberman Foundation administers its Star Teacher Pre-Screener assessment to determine the match of

[*] See http://www.mo.nesinc.com/
[†] See http://www.teachforamerica.org/why-teach-for-america/who-we-look-for

candidates for teaching in urban classrooms.[*] And many other teacher preparation programs administer some sort of assessment of candidate attitudes, dispositions, or relevant prior experiences as an admissions screen or upon program admission. However, such locally developed assessments are often not empirically validated instruments and would likely not provide the basis for measurable comparisons between programs.

The objective for using this indicator for program improvement purposes would be two-fold:

1. To provide programs with good information about their students' habits, attitudes, and prior experiences that might inform decisions about appropriate coursework or support for each teacher candidate;

2. To discover whether there are differences between candidates with different teaching promise profiles with respect to their success in the program or in the classroom as teachers.

Used in the context of program accountability, a valid and reliable assessment of teaching promise would provide an indication to state officials and the public of how selective programs are in choosing candidates most likely to succeed in the teaching profession and (importantly for potential enrollees) perhaps what factors are most important to individual programs.

Candidate/Completer Diversity—Diversifying the teacher workforce is a widely shared goal among educators and state officials, and there are a number of research studies that show the benefits—particularly for minority students—of having teachers who share their life experiences.[†] The great majority of states, however, are far from having a teacher corps that mirrors the increasing diversity of their students; minority representation in the entire U.S. teacher corps is 18%, while the percentage of minority students in our schools is 49% and steadily increasing.[‡] Moreover, it is not realistic to expect teacher preparation programs to enroll—and graduate—a significant percentage of minority candidates if the college or university in which they are situated doesn't. And given the academic and social challenges that many minorities face upon enrolling in an institution of higher education, graduation and program completion rates for minority teacher candidates may be much lower than those of other candidates even when a significant percentage of program entrants are minorities.

These realities have several implications for crafting measures of candidate diversity. First, programs need to track the program performance, completion rates, and—insofar as possible—

[*] See http://www.habermanfoundation.org/starteacherprescreener.aspx

[†] See, for example, M.C. Eddy and D. Easton-Brooks, D. (2011). Ethnic Matching, School Placement, and Mathematics Achievement of African American Students from Kindergarten through Fifth Grade. *Urban Education*, (46), 1280-1299

[‡] Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service. (2016). *The State of Racial Diversity in the Educator Workforce,* Washington, D.C.: U.S. Department of Education

the classroom teaching success of all sub-groups of candidates by race/ethnicity just as they are encouraged to do with respect to all candidates' academic strength. If there are disparities among sub-groups with respect to program and teaching outcomes, there is good reason to believe that the program is simply not serving the needs of low-performing sub-groups. Second, although it is unrealistic to expect that a teacher preparation program's percentage of minority candidates will significantly exceed the percentage of minority enrollees at the host institution, it is reasonable to expect the EPP to undertake efforts to ensure that the percentage of minority students in teacher education at least mirrors the percentage of minority students enrolled at the university.

Since individual certification programs have relatively little leverage to recruit teacher candidates to a specific teaching field, the relationship between the percentage of candidates in teacher education and the larger university is suggested as an EPP-wide indicator. A particular EPP may have few minority candidates in some programs, but for all of its preparation programs taken together minority enrollment (and completion rates) should be representative of the university population. The completion rate of sub-groups of candidates is seen here as a program-specific indicator, however, because it seems reasonable to hold individual programs accountable for ensuring that they can meet the needs of all their admitted candidates and can ensure that struggling candidates have access to appropriate support and any necessary remediation.

Because increasing the participation of minority candidates in teaching is such a compelling interest in Illinois, displaying the number of minority completers from the various programs together with the percentages will provide a clear indication of how many minority completers programs are contributing to the teacher pipeline.

Mastery of Teaching Subjects—The assurance that candidates who complete teacher preparation programs have broad and deep knowledge of the subjects they're certified to teach is one of the most important responsibilities of the programs. This responsibility applies even if candidates don't acquire their content knowledge in the teacher preparation program per se—as is the case with most secondary subject fields and post-baccalaureate programs. Programs can enforce high content knowledge standards by requiring completers to have a solid college GPA in their subject courses, recommend them for licensure contingent upon their receiving a strong score on the relevant state licensure examination, and admit post-baccalaureate candidates only if they score well on the appropriate GRE.

For purposes of program improvement, undergraduate programs have a clear advantage over post-baccalaureate programs because teacher education faculty can work with the arts and sciences faculty to help ensure that the courses required for the major serve the needs of teacher candidates. And candidate program outcomes on assessments of content knowledge

can provide evidence to the arts and sciences faculty that their content courses may or may not be meeting candidate needs.

The suggested measures of candidate content knowledge are strictly program-specific, and they apply—as is the case for all the summative assessments of candidate knowledge and skill—to the most recent cohort of program completers. The measures may include—as required by the U.S. Department of Education and many state education agencies—the pass rate of candidates on state licensure exams. However, pass rates are information-poor and often rely on a relatively low bar—a bar that can differ significantly between states that use the same content knowledge examination (e.g., the Praxis II series).

Where content knowledge test developers or other experts can identify a specific assessment score that represents a strong command of subject knowledge, measures of the proximity of program candidates' scores to that score are far preferable to pass rates. Where that is not the case, states may set the desired assessment score as a norm-referenced value, such as the 50th *percentile* or 60th percentile in the statewide (or, if known, national) scoring distribution. In any case, program performance measures would be calculated by measuring the percentage of candidates who score at or above that value or by measuring the extent to which the average score for a candidate cohort exceeds, meets, or falls short of that value.

On the Illinois Content Area Tests, the passing score for all subjects is set at 240. If this is a score that denotes considerable command a candidate's teaching subject(s), then it is an appropriate benchmark to help drive program excellence. If, however, it denotes mere competency and is reached by the overwhelming majority of teacher candidates, continuous program improvement may be better served by aiming at a higher score that denotes mastery in teaching skills, not merely proficiency.

In addition to using the *mean* cohort score as a performance measure, it is important to convey the distribution of scores in a cohort. A mean score for a cohort gives no indication of whether a significant number of candidates had very high scores or very low scores that all balanced out in the mean score. If the goal of a teacher preparation program is to graduate teachers who will be effective in the classroom, it is particularly important to know whether a significant number of those graduates have weak skills whatever the average for all graduates might be. Such distribution measures could be stated, for example, as the percentage of candidates who scored below the 33rd percentile or above the 66th percentile in the statewide scoring distribution. In a *criterion-referenced* context, it could be stated as the percentage of candidates who scored at several meaningful benchmarks that denote different levels of competence (e.g., inadequate, proficient, superior). Other measures of distribution are also possible, such as the use of *Standard Deviation*.

These same measurement options apply in the case of all other recommended indicators that involve assessments for which specific benchmarks have been identified that denote proficient or exemplary performance.

Subject-Specific Pedagogical Knowledge—This indicator refers to the specific kind of knowledge of a subject field that enable a teacher to teach it effectively. It is sometimes referred to as "Pedagogical Content Knowledge" or "Content Knowledge for Teaching," and it requires a good understanding of how students learn the key content of a subject and the greatest challenges to mastering it. Having this knowledge enables a teacher, for example, present key concepts of a subject matter in a manner that will promote deep understanding. Or, it enables her to look at a mistaken answer a student gives in a mathematical operation and know exactly what the student's thinking process was in order to have made that particular mistake.[*]

In a standard teacher preparation program, this kind of knowledge may be acquired by candidates in a single Methods course. For the most part, however, programs fail to provide adequate instruction in subject-specific pedagogical knowledge. If a number of eminent teacher educators (Deborah Ball, Lee Shulman, and others) are correct, this is a major failing because, according their teaching practice and empirical research, subject-specific pedagogical knowledge is vital for effective teaching.

Not only is this kind of knowledge barely taught, but it is largely not assessed. It is not tested in content knowledge assessments for licensure or in traditional pen-and-paper licensure tests of teaching skill, such as the Illinois Test of Academic Proficiency (TAP) or the Principles of Learning and Teaching (PLT). To some extent, the edTPA and the new Praxis Performance Assessment for Teachers (PPAT) do assess this kind of knowledge, but only for a narrow span of a candidate's teaching subject(s).

There is one broader and deeper assessment of subject-specific pedagogical knowledge that has recently come on the market, the Content Knowledge for Teaching (CKT) assessment by ETS.[†] At present, however, it is available only for elementary school teachers, and so the use of this indicator for both program improvement and program accountability purposes at this point remains largely aspirational—at least for the moment.

General Teaching Skill—Adequate teaching proficiency—the fundamental teaching skills required to teach effectively in any subject—is another key outcome that teacher preparation programs need to verify as having been demonstrated by all program completers. Although preparation programs individually generally administer several assessments of candidates' teaching skill, until recently there was no widely adopted summative program assessment of

[*] See D. L. Ball, H.C. Hill, H.C, and H. (2005). Knowing Mathematics for Teaching: Who Knows Mathematics Well Enough to Teach Third Grade, and How Can We Decide? *American Educator*, 29(1), 14-17, 20-22, 43-46.
[†] Information on this assessment can be found at https://www.ets.org/s/educator_licensure/ckt_handout.pdf

candidates' skill that required a true demonstration of teaching and could be nationally normed so that completers in all programs using the assessment could be compared. The only nationally normed summative assessments of teaching skill were paper-and-pencil tests (such as the PLT or TPA), which many states still use as a licensure assessment but which many educators believe is not a sufficiently "authentic" assessment of actual teaching proficiency.

The edTPA has changed the picture, and it and a similar, more recent assessment that has come onto the market (the PPAT) now make it possible to administer assessments of actual teaching skill involving observations of candidates' teaching sessions and to compare the scores of test-takers at least statewide. A few states, including Illinois, use the edTPA as a requirement for teacher licensure, and in a growing number of states it has completely supplanted the paper-and-pencil assessments of teaching skill.

For purposes of program improvement, it may be helpful to give programs candidate performance data on each of the specific teaching skill assessment domains. For accountability and public information purposes, however, the overall performance of candidates in individual programs should be satisfactory. For the edTPA, as for other licensure assessments, the pass rate measure (currently required by ISBE) is not a robust measure of program performance. A better measure, similar to the recommendation for the content knowledge assessment, would be to provide the mean scores and score distribution of program candidate cohorts and compare the mean scores to the statewide (or regional or national) mean or to an acknowledged and verified proficiency benchmark.

In 2014, the Illinois State Board of Education worked with teacher preparation programs and national experts to determine a passing score that will roll out during the next four years, beginning at 35 out of 75 in Sept. 2015 and climbing to 41 by Sept. 2019. Some states have higher passing scores, however, and there are also state mastery scores that are even higher. Illinois may find it helpful to revisit its edTPA benchmarks during the pilot phase of system development. As in the case of the state's content knowledge tests, a higher benchmark for mastery may better serve continuous program improvement.

New Completer Rating of Program—It seems quite appropriate to want to find out how satisfied candidates and completers are with the programs that prepared them. Such consumer feedback ought to be valuable to efforts of program staff and faculty to improve the program, and it should provide the public with a sense of greater transparency about program performance and greater confidence that programs will be responsive to the needs of their candidates. From the standpoint of both program improvement and accountability, however, the challenge is obtaining responses from candidates and completers that are reliable and useful and that can provide a basis for meaningful comparison between the performance of different preparation programs. This requires the use of the same, well-developed survey

instrument for all preparation programs statewide and a good response rate from candidates. Both these conditions are easy to satisfy in surveys of recent completers (or late-stage candidates) because they are readily reachable and can be compelled to complete such a survey as a condition for program completion or licensure.

However, states that conduct such a survey (and approximately 15 states do) have often found—in an accountability context—that the responses of recent completers often don't provide significant differentiation between one preparation program and another. And teacher educators have found that the information they receive from recent completers, who have not yet had significant experience trying to apply what they learned in their preparation program to their work as full-time teachers in a classroom, is not as helpful as the information they receive from completers who have been teaching for a year or more.

Because there are no established benchmarks for program performance on completer satisfaction surveys, the measures suggested for this indicator are simply norm-referenced scores—i.e., how programs perform compared to the statewide mean for scores on the surveys. It may be valuable to compare performance scores of programs in all certification areas (e.g., secondary science, middle school math, etc.)—not just within each certification area—but that may require standardization of scores within each area since candidates in some program certification areas may tend to be more critical or less critical than candidates in other areas.

Novice Teacher Rating of Program—A number of states have found that the perceptions of completers with substantive teaching experience about the strengths and weaknesses of their preparation programs are significantly more valuable than the perceptions of recent completers who lack that experience. Because of this, some of these states administer satisfaction surveys only to novice teachers. The biggest challenge in administering surveys to novice teachers, however, is ensuring a good response rate. It is much more difficult to track down program completers once they've left the programs, and a substantial number of completers may have moved to other states. States and preparation programs also have less leverage over novice teachers than they do over late-stage candidates or recent completers to compel their response to the surveys. Some states have addressed the leverage problem by requiring novice teachers to complete program satisfaction surveys as a condition for receiving second-stage licensure, but not all states have a tiered licensure system. Illinois has a second stage licensure after four years in the classroom, but that is likely too distant to tie that second stage license to the completion of a program survey in a teacher's first year.

In any case, comments from novice teachers about the adequacy of their preparation program are likely to be of great interest to programs for improvement purposes. If it is difficult to ensure a strong and representative response of those teachers to a survey about their

perceptions of their programs, however, the value of such surveys for accountability purposes is somewhat compromised. This does not mean that these surveys can't be used for accountability and for comparisons between programs, but it may mean that only survey results at the upper and lower limits of the statewide score distribution should be regarded as noteworthy from an accountability standpoint.

Principal/Supervisor Rating of Program—There are about as many states that use surveys of novice teachers' principals or supervisors about the effectiveness of their new teachers' preparation as use surveys of candidates and completers. These surveys also serve to ensure the public that programs are accountable for being responsive to consumers—in this case the individuals who hire the teachers the programs produce. And they provide a direct opportunity for K-12 officials to weigh in on the adequacy of preparation programs and potentially to influence change in program practices and policies.

There are, however, challenges facing the use of this indicator. States have very little leverage to compel principals or their surrogates to complete these surveys. Principals often have little time to devote to observations of new teachers or to conferences with them, and even when they do principals may not be in the best position to judge how effective new teachers really are. Others in supervisory roles may be better positioned to provide the requested data, but in either case there is no guarantee that the ratings will be consistent between supervisors—even in the same school. Moreover, although this indicator is intended to provide ratings of how well prepared new teachers were by their preparation programs, it is difficult in actual practice to distinguish between a rating of preparation and a rating of teacher's competence.

For these reasons, TPA does not advocate for the administration of such a survey of principals and supervisors. Instead, TPA believes that a validated and well-administered classroom observation protocol of new teachers provides far more valid and reliable information.

Impact on K-12 Students—Many researchers and educators regard the learning gains of a teacher's students to be the most compelling evidence of the teacher's success. All other outcomes are a proxy for student learning. There is, of course, a good deal of controversy concerning what constitutes evidence of meaningful learning. Opponents of the kind of standardized testing required for *value-added assessment systems* and many *student growth models*, argue that such tests do not assess meaningful learning and that they disadvantage students who are simply not good test-takers.

Standardized testing has been used for generations, however, to provide universal measures of student academic proficiency that enable valid comparisons between all children. And as the student learning standards movement has taken hold in virtually all states, there has been an increasing effort to ensure that both statewide learning assessments and the curriculum of tested subjects are aligned with those standards. This alignment makes it that much more likely

that teachers are teaching the material children in the state will be tested on and that much more plausible to hold teachers responsible for how their students perform on the assessments.

Thus, many states, including Illinois, have adopted an annual teacher performance assessment that includes, at least in part, an assessment of teacher impact based on their students test scores or learning growth. Such learning outcomes-based evaluations of teachers are highly controversial, and there are highly technical discussions and disagreements among experts about the comparative reliability and validity of the various kinds of statistical models. There are also frequent cautions from education researchers that because these models may in fact attribute inappropriate scores to teachers in isolated cases, they should not be used as the sole basis of evidence for high-stakes decisions about teacher dismissal, promotion, or compensation.

One of the interesting considerations in using such teacher impact data as evidence of the performance of the teacher education programs that prepared them is that erroneous scores that may be derived from the statistical models in isolated cases become virtually insignificant when it is the average score of multiple cases that is the metric. This is precisely the situation for the Teacher Impact indicator as a cohort measure. Many people have concerns that specific school factors can be "wild cards" in the outcomes that determine teachers' impact scores. But if those factors are controlled for statistically—and especially if teachers in similar school situations are compared to one another as suggested in the recommended measures—it is that much more likely that significant differences in aggregated cohort impact scores are attributable to differences between programs.

The suggested cohort to be measured on this indicator, as for all others indicators under Completer Proficiency as Teachers, is the combination of the second most recent and third most recent cohort of program completers—i.e., those completers who are likely to have spent 1-2 years in the classroom. There are two main reasons for this: to increase the size of the measured cohort for the sake of *statistical power* (especially important in the case of small programs) and to draw on multiple years of performance data for at least some candidates to balance out potential anomalies in a single year of their teaching performance. Some states follow completer cohorts as much as five years out, but after two years of full-time teaching it seems increasingly difficult to separate out the effects of the preparation program on a teacher's performance from the effects of experience and continuing professional development.[*] Also, the more distant in time the cohort is from the present, the more the

[*] For a discussion of the waning influence of preparation program impact on completer performance over time, see G. Henry, C. Thompson, K. Fortner, K. Purtell, R. Zulli, and D. Kershaw, D. (2010). *The Impact of Teacher Preparation on Student Learning in North Carolina Public Schools*. Technical Report. Chapel Hill: Carolina Institute

program may have evolved and the less relevant completers' performance may be to present program improvement needs.

Teacher impact and all other measures of program completers as teachers in the classroom face the challenge of tracking individuals who may have left the state or who, as faculty in private schools, are likely to be exempt from annual performance evaluations that include student impact scores. That will be a problem for some programs more than others, and the information nevertheless should be valuable for program improvement. But in an accountability context, state officials will have to decide how to handle those measures where *sample attrition* is high. It may be prudent to focus specifically on programs that are either very high-scoring or very low-scoring on these measures.

Using such teacher impact data, several states (Louisiana, Tennessee, North Carolina) have found that large differences in the mean teacher impact scores between programs in the same certification field can indeed be indicative of important differences in program structure or policies. Changes then made in the lower-scoring programs to emulate the higher-scoring programs have had positive results for their subsequent teacher impact scores and other outcomes. It is this utilitarian promise in using teacher impact data—quite apart from the controversies over their adequacy and validity—that is the principal reason for including such data in the program improvement and accountability system.

Not all states, however, have found teacher impact data to be useful for accountability purposes. In piloting their own growth model for the purposes of distinguishing higher-performing and lower-performing programs, researchers in Missouri found that there was much more variation in student growth scores among teachers within individual teacher preparation programs at an EPP (in this case Elementary Education) than there was between like programs at different EPPs.* From the standpoint of program improvement, however, even this outcome is instructive; it is a reminder that aggregated scores do not provide the whole picture and that significant disparities in candidate performance within programs should be a significant driver of program change. Indeed, the teacher impact measures recommended for the Illinois program improvement and accountability system include distribution statistics in addition to mean scores.

It will be a challenge for Illinois to implement this indicator because it will require, first, separating out the student growth component of teachers' annual performance evaluation from the other components. And, second, it will require some assurance that the student

for Public Policy. These authors suggest that program influence is virtually nil by the fifth year after program completion.

* See C. Koedel, E. Parsons, M. Podgursky, and M. Ehlert. (2012). Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs? CALDER Working Paper No. 79. Washington, DC: American Institutes for Research.

growth data can be standardized for all districts in the state. This should be easier to accomplish in state-tested subjects, but it is likely to be very challenging in subjects that rely on locally developed assessments as the basis of student growth data.

Some states, facing the similar challenge of separating out scores on different components of a combined teacher performance assessment have opted simply to use the combined teacher performance score as the basis for a preparation program performance measure. There are several drawbacks to such an approach, however. The combined score of a multi-part assessment does not by itself provide much actionable information to aid program improvement efforts. Moreover, when scores of different components are combined into a single score, it becomes an aggregated score that communicates low-value information and tends to reveal much less variation in performance. A second drawback is that to the extent the combined score is based on district-initiated assessments and district-specific ratings of performance, teachers' scores will not be truly comparable between districts. Third—and this could be a problem for using individual performance assessment components, as well—contracts with state or district teacher associations may prohibit the use of annual teacher performance assessment data for any external purpose.

Demonstrated Teaching Skill—How program completers perform as teachers of record in their own classroom is another compelling outcome that—at least in the early stages of teachers' professional careers—would seem to be attributable in part to the programs that prepared them. And although there are disagreements about the validity and reliability of different protocols, classroom observations that used appropriately trained observers and well-tested protocols are widely accepted by educators as important and useful assessments of teaching skill and are less controversial than attempts to assess teacher impact on student learning.

If this indicator is to be implemented statewide, however, and especially to be used as an accountability construct to determine and compare preparation program performance, the same observation protocol (or compatible protocols) with comparably trained observers must be employed statewide (at least for each individual certification level, e.g., early education, elementary education, secondary education). This is currently not the case statewide in Illinois, where observations of teachers are conducted as another component of their annual performance evaluation. Not only would classroom observation scores need to be separated out from the combined performance score (as in the case of student growth scores), but there would need to be standardization of the observation *protocol* scores so that scores under different protocols can be compared. This may be possible with the Danielson and Marzano observation protocols, which are widely used in Illinois but are apparently not used in every district.

If classroom observation performance scores can be standardized for each observation protocol around comparable benchmarks (e.g., proficient, excellent, etc.), then the program performance scores could compare the program cohort mean scores to the benchmark scores. The suggested measures also include distribution statistics, which could be the percentage of completers scoring below the proficient level or at the excellent level or above. If the protocol benchmarks are not acceptable to ISBE and other stakeholders, program scores could be based on statewide mean scores and distribution scores. Programs could be scored, for example, on the percentage of completers scoring at or above the state mean, above the 66th percentile, and below the 33rd percentile.

K-12 Student Perceptions of Teachers—There is increasing interest around the U.S. in using appropriately constructed and administered surveys of K-12 students' perceptions of their teachers' effectiveness as a source of important information about teachers' classroom performance. Only a very few states have expressed an intention to use these data to assess the performance of teacher preparation programs, but one such survey (Tripod) was used in the Gates Foundation-funded MET project that studied the validity of value-added data, classroom observation data, and K-12 student survey data for evaluation teachers.[*] The study corroborated the validity and utility of all three kinds of teacher assessment and also found that, when used together (any two or all three), the predictive validity of the combined assessments exceeded that of each assessment individually.

ISBE and the Steering Committee expressed interest in exploring the possibility of adding K-12 student perceptions as an indicator for preparation program improvement and/or accountability down the road, but it is not likely to be included in the initial rollout of the new program improvement and accountability system.

Entry into Teaching—In the interest of preparation program efficiency and the state's ability to identify and respond to its need for teachers generally, for specific subjects, and for specific school districts, the higher the percentage of program completers who enter teaching the better. Some programs see larger percentages of their completers take jobs out of state, and it will be more difficult to determine truly accurate teacher placement rates in such cases. But, taking into account such circumstances, if preparation programs have relatively low rates of completer entry into the profession, that may signal a serious problem that needs to be investigated further.

Since recent program completers may delay their entry into teaching positions for a variety of reasons, many states define placement rates over two or even three school years out from completion. The recommendation here for the Illinois system is to follow placement over two

[*] S. Cantrell and T. Kane. (2013). *Ensuring Fair and Reliable Measures of Effective Teaching*. Seattle, WA: Bill & Melinda Gates Foundation.

years because state officials generally find that relatively few completers take their first teaching position later than that. Thus, the suggested measured cohort for this indicator is the combination of the most recent and second most recent program completer cohorts. [In the fall of 2016, for example, the most recent cohort would be the 2016 completers who may just now have entered teaching and the completers who finished the program in 2015.]

There is no defined benchmark for what rate of placement teacher preparation programs should be expected to meet, and partly because of the difficulty of tracking program completers any national percentage for teacher placement is likely to be inaccurate. Moreover, there are significant differences in the demand for teachers depending upon their teaching subject. Though teaching and nursing are far from fully comparable professions, it may be of some value in setting a placement rate benchmark for teaching to note that placement for nursing school graduates after six months is reported to be 86-95% depending upon geographic region and nursing credentials.[*] For the short term, until there is greater clarity about the completer placement rate of all preparation programs in Illinois and a performance benchmark can be set, the suggested measure is simply to report the placement rate for each program and then compare it to the statewide program mean (ideally, both the mean for all programs and the mean for each individual certification field).

Completer entry percentages are expected to be reported out by individual program, as well as EPP-wide. Equally important would be reporting out entry percentages for high-need subjects, which may be taught by completers from several different programs within the same EPP.

Some states only include in their job placement statistics completers who are teaching in their field of certification. The recommendation here, however, is to include all completers who are hired into teaching positions because preparation programs have no role in making school assignments and cannot fairly be held accountable for out-of-field placements. Also recommended is the inclusion of completers who take jobs in other education roles, although it may prove too difficult to track them.

Persistence in Teaching—From a teacher supply and demand standpoint, it is clearly important to track the persistence of teachers once they enter the profession. From an accountability standpoint, it may seem less apparent that preparation programs should be held accountable for how long their completers remain in teaching. State officials concerned with teacher supply and district officials who are responsible for hiring teachers would certainly want to know, however, whether there are significant differences in the persistence of completers from different programs. Such differences may be largely a function of economic circumstances—the kinds of schools and districts in which completers are placed (i.e., whether they historically

[*] American Association of Colleges of Nursing. (2015, October). Research Brief. Downloaded 9-19-16 from http://www.aacn.nche.edu/leading-initiatives/research-data/Research-Brief-2015.pdf

have high turnover) or the production and placement of teachers in subjects and regions that have a glut of teachers available (though that would more likely be picked up in teacher entry measures).

However, there are reasonable expectations that persistence rates can be affected by program practices, and there is research to support those expectations.[*] Programs that specifically prepare teachers to teach in high-needs schools, for example, arguably should have much lower turnover rates for their completers in those schools than programs lacking that emphasis. And regardless of the specific mission of teacher preparation programs, it entirely possible that programs with very low turnover rates simply do a better job of preparing their teachers with the knowledge and skills they need to stick it out or do a better job of screening out prospective candidates who are less clearly committed to teaching or who lack the self-confidence and determination required to cope with some of the daunting challenges teaching can present. These are testable hypotheses, and an investigation into significant disparities between programs in their rates of teacher retention may be able to confirm or refute them. Especially at this early stage in the implementation of the Illinois program improvement and accountability system, this is one of the express purposes for collecting all the data recommended.

Some states define persistence as retention in an individual school or district, but most states define it as remaining in the teaching profession or in an education role (e.g., principal). As with entry into the profession, there is no easily determined benchmark for teacher persistence. Many articles have been written about the high rate of turnover in the teaching profession, which turns out not to be significantly higher than the turnover rate in many other professions.[†] Most educators would agree that a higher persistence rate ideally is better than a low persistence rate. However, persistence rates are also a function of social attitudes and norms, and we are told that we live at a time when there is greater churn in the job market generally and more changing of jobs than at any time in our nation's history.[‡]

In the absence of a clear benchmark for persistence, it would be helpful for both state officials and preparation programs to have long-term trend data on teacher retention that could point not only to overall trends but also to significant changes in the retention rates for individual programs. With these data in-hand, state officials and teacher educators could attempt to

---

[*] See R. Ingersoll, L. Merrill, and H. May. (2014). *What Are the Effects of Teacher Education and Preparation on Beginning Teacher Attrition?* CPRE Research Report #RR-82. Philadelphia: Consortium for Policy Research in Education.

[†] D.N. Harris and S.J. Adams. (2007). Understanding the Level and Causes of Teacher Attrition: A Comparison with Other Professions. *Economics of Education Review* 26 (3), 325-337. See also Ingersoll, 2001.

[‡] See, for example, Bureau of Labor Statistics. (2015, March 31). Number of Jobs Held, Labor Market Activity, and Earnings Growth among the Youngest Baby Boomers: Results from a Longitudinal Survey. News Release. Washington, DC: U.S. Department of Labor.

determine what accounts for any notable shifts in retention rates over time—perhaps changes in program practices, in the characteristics of admitted program candidates, in the kinds of schools in which completers are placed, in district professional development practices, in the regional economy, etc. However, the problem with historical trend data even only five years old is that the data will reflect conditions and practices five years ago that may not characterize the present realities of either programs or K-12 schools or the economy. From a program improvement standpoint, five-year-old data are likely to be only minimally useful. And from an accountability standpoint, it is likely to be neither productive nor appropriate to base any decisions about present program adequacy on the program's performance five years ago.

Thus, the suggestion here for the Illinois preparation program improvement and accountability system is to measure program performance on completer persistence using four years of post-completion data. Specifically, the measure would be the percentage of completers from the fourth most recent completer cohort who persist in teaching (after their initial entry) for one, two, and three years. [In the fall of 2016, this would be the cohort that completed the program in 2013, and its members will have persisted in teaching for up to four years if they entered teaching in 2013 and are still teaching in the 2016-17 academic year.]

This specific measure has the advantage of using relatively recent data so that meaningful program improvement and accountability measures arguably can be based on those data. They have the disadvantage of not conveying a longer-term completer retention rate, but completer retention after a few years is likely to be much more a function of factors other than the preparation program itself. What should be particularly instructive for both state officials and teacher educators is to undertake two basic comparisons:

1. A comparison of persistence rates that are especially high and especially low among all programs
2. Eventually, a comparison of current persistence rates of all programs with historical rates in order to discern significant trends or deviations.

As in the case of Entry into Teaching, persistence of teachers who are certified in high-need subjects can be tracked separately for each EPP.

Illinois state officials can make an important contribution to the decision about the appropriate measures for teacher persistence. If they have access to historical persistence data by program (or even by EPP), they should be able to determine whether there is a significant difference between two-year, three-year, and four-year persistence rates and how many years after entry into teaching the greatest fall-off in persistence tends to occur. They also might find interesting and important persistence trends—perhaps, for example, that teachers who remain in teaching for at least four years have a strong likelihood of remaining for at least four more. Such information is not directly relevant to preparation program improvement and accountability,

but it could spur the exploration of ways that EPPs, districts, and state agencies could work together to improve teacher persistence.

Placement in High-Needs Schools—Both state and district education officials have a particular concern with placing (and retaining) teachers for *high-needs schools* that tend to have high teacher turnover and often are not able to offer classes, at all, in subjects that many schools in the state find difficult to staff. Typically, the most difficult placements are in inner-city urban schools and remote rural schools. Not every program is going to send significant numbers of completers to such schools; typically—especially outside of urban areas—completers take jobs in proximity to the location of their preparation program. It will be helpful to state and district officials to confirm the extent to which programs place completers in such schools and perhaps incentivize increased placement from the feeder programs and others.

The suggested measures for this indicator are parallel to those for entry into teaching, tracking placement of the most recent and second most recent program completer cohort members for two years.

Persistence in High-Needs Schools—Given the frequently high turnover rate in high-needs schools, persistence of teachers in those environments is especially important both for organizational stability and the development of capable and experience teachers. Because of the challenging nature of such schools—and sometimes simply because of the remoteness of their location—teacher retention is itself a challenging proposition. As with persistence in the profession in general, however, if there are significant differences in the persistence rates in high-needs schools of completers from different programs, this may point to important program strengths and weaknesses that need to be investigated.

The suggested measures for this indicator are similar to those for the more general persistence-in-the profession indicator: the percentage of completers from the third most recent completer cohort who are teaching in high-needs schools one, two, and three years after program completion.

Completers in High-Need Subjects—Because it is important for state and district officials to know the supply and sources of teachers in *high-need subjects*, this indicator is recommended for the Illinois system. It may not be the specific mission of all EPPs to offer programs in every high-need subject, but it seems reasonable to expect that especially state-supported EPPs will take on some share of the responsibility for producing teachers for critical shortage areas. Some teacher educators and other experts believe it is the placement rate of completers in high-need subjects that is important, but the recommendation here is to focus on all completers, independent of actual job placement, because they constitute (along with unemployed certified teachers) the potential supply for those subjects.

Thus, the suggested measures for this indicator are the number and percentage of program completers in each high-need subject. This may require combining completers from different individual programs that offer different certifications but all entitle the recipients to teach the same high-need subject or subjects. The percentages for each EPP—and possibly for each high-need subject in that EPP—can then be compared with the statewide *median* since there will be so much variation between EPPs that the mean score is meaningless as a point of comparison.

Minority Completers—Increasing the supply of minority teachers is a major national priority and an expressed priority of the PEP Steering Committee. Illinois has not only a substantial African-American in its larger cities but a growing Hispanic population in both its urban and rural areas. The Candidate/Completer Diversity indicator already requires the collection of data on minority program completers by sub-group in order to determine the proportion of entering candidates who reach program completion. The completion data, disaggregated by sub-group at the EPP level as the total number and percentage of minority completers among all completers, can be reported for the Minority Completer indicator. Although not all EPPs can be expected to produce large numbers or percentages of minority teachers, it should still prove useful to state and district officials to compare the data for each EPP to all others statewide (perhaps to the statewide median.

NEXT STEPS FOR THE ILLINOIS PROGRAM IMPROVEMENT SYSTEM

In its six months of work, the PEP Steering Committee succeeded in reaching a general consensus on the importance or potential importance of 18 indicators and a number of corresponding measures of programs' performance on those indicators. This set of indicators and measures remains preliminary, however. The utility and reliability of the measures, as well as the quality and availability of the data necessary to populate them will be examined in a two-year pilot prior to actual implementation of the program improvement and accountability system.

There are additional components of the annual program review process that neither ISBE nor the Steering Committee has yet had sufficient opportunity to recommend for adoption in the program improvement and accountability process. These elements, summarized below, also will be proposed, discussed, and tested in the pilot process:

1. *Program Performance Targets and Thresholds*—It will be necessary for ISBE and PEP to choose the benchmarks for every measure that distinguish between satisfactory and unsatisfactory performance and denote exemplary performance. If there is high confidence in the validity and reliability of a measure, the difference between satisfactory and unsatisfactory performance may be determined solely based on

programs' performance scores on the measure. Where confidence in the validity and reliability of the measure is lower, it may be necessary to consider additional information in determining whether a program's score on a measure is a true indication of a program deficiency requiring further attention. And in deciding whether a program's performance is truly exemplary, TPA recommends that not only the score on the performance measure be considered but that an examination of related program practices be undertaken, as well, to determine whether there are specific program practices that account for the high score and that other programs might emulate.

2. *Significance of Scores on Individual Measures*—ISBE and PEP will have to determine whether to regard each individual performance score in its own right, whether some measures should be given more attention than others, and whether the individual measures should be combined into a single overall program score or rating. If there are individual measures that are considered significantly less important than others and the consensus is either generally to ignore low scores on any such measures or give them minimum weight in an overall program score, it is important to rethink whether these measures truly belong in the program accountability system. If an overall program score is a desired feature of the program improvement and accountability system, it is important to ensure that any variable weighting of the measures has a firm, defensible basis and that assigning an overall score doesn't preclude disclosing any considerable variations in a program's performance on individual measures.

3. *Consequences of low program performance*—This is a consideration involving both strategy and institutional capacity. If the ultimate goal of the system is program improvement, what kind of response to the performance data will best achieve that end? Are both ISBE and individual EPPs prepared to respond to low program scores on every performance measures? Or will the focus of attention be on programs that have multiple low scores on measures or a comparatively low overall program score? Will programs with low performance scores be placed on some sort of "watch list" in the expectation that their low scores will improve within a given period of time? What will the role of ISBE be (if any), in assisting the efforts of EPPs to address programs' low performance? What will be the consequences for programs that prove unable to raise their performance measures over a reasonable period of time?

4. *Public reporting of program performance*—It is a reasonable expectation, consistent with the accountability of ISBE to the public, that information on the performance of preparation programs will be made public. This requires a decision about several considerations, including (a) whether performance on all measures or only select

measures will be shared publicly and (b) how the performance information will be disseminated so that it is both meaningful and understandable to a lay audience and—if simplified or aggregated in any way—will nevertheless convey an accurate picture of programs' performance.

5. *Periodic review of the indicators and measures* – As with all other professions, the knowledge and skills required of teachers evolve over time. Thus, the knowledge and skills to be taught in professional preparation programs must also evolve. And it is only natural to assume that way we assess professional preparation must also change over time. We've noted earlier in this Guide that adequate implementation of some of the important indicators of preparation program performance is hindered by the absence of adequate assessments and that poor-quality data hamper their implementation, as well. Thus, it seems altogether appropriate and necessary that every five years or so ISBE assemble a group similar in composition to the PEP Steering Committee to review what has been learned from the current accountability system and to consider ways to improve upon it. Continuous improvement of our educator preparation program goes hand in hand with continuous improvement in our means of assessing their performance and in our ability to provide guidance for needed change.

GLOSSARY OF TECHNICAL TERMS

Candidate—an individual who is enrolled in a preparation program to become a classroom teacher.

Certification—recognition by state authorities that an individual has fulfilled the requirements to teach one or more specific subjects in specific grades in the state's public schools. Sometimes called "endorsement" or "licensure".

Cohort—the group of teacher candidates or completers whose performance or status is the basis of measurement.

Completer—an individual who has successfully completed a teacher preparation programs. Often called a "graduate".

Criterion-referenced measures—measures based on a specific score (or scores) that denotes an important performance level (passing, proficient, outstanding, etc.) on a fixed set of criteria. In criterion-based measurement, everyone or no one could possibly achieve a performance benchmark.

edTPA—a specific assessment of late-stage teacher candidates that uses video clips of candidates' classroom teaching, evidence of promoting student learning, and other artifacts to evaluate candidates teaching skill in his/her teaching field. In some states, passing this assessment is required for certification.

Education program provider—the college, school, department, or organization that offers one or more programs or pathways leading to teacher certification in specific fields.

EPP—acronym for Education Program Provider.

GPA—acronym for Grade Point Average.

High-need subject—defined by the U.S. Department of Education as bilingual education and English language acquisition, foreign language, mathematics, reading specialist, science, and special education, plus any other field identified as high-need by a state government or local education agency.

High-needs school—defined by the U.S. Department of Education as a school ranking in the top quartile nationally for the number of unfilled teaching positions or located in an area where at least 30% of students live in poverty.

Indicator—outcome or characteristic of a preparation program, program provider, candidate or completer than can be measured to gauge program performance. Also called a "performance indicator".

ISBE—Illinois State Board of Education, the state agency responsible for teacher preparation program approval and review.

Licensure—recognition by state authorities that an individual has fulfilled the requirements to be a teacher in the state's public schools. States usually issue different licenses for beginning and more experienced teachers.

Mean—average, which is obtained by calculating the sum of the scores of members of a group and dividing that total by the number of members.

Measure—quantity or rating on an assessment or other data source that provides the basic data for determining the performance of a program, completer, or candidate on an indicator. Also called a "performance measure".

Median—number or score that is the mid-point among all numbers or scores under consideration. Out of five different numbers arrayed in numerical order, for example, the median would be the third number.

Norm-referenced measure—measures denoting individuals' placement on a distribution of scores from the highest to the lowest, often expressed as a percentile rank. Whether or not individuals meet a certain set of performance criteria is irrelevant in norm-referenced measurement.

Pathway—any course of candidate study or prospective teacher assessment that culminates in certification as a teacher upon successful completion.

PEP—acronym for Partnership for Education Progress, the effort funded by the Joyce Foundation to develop the Illinois Teacher Preparation Program Improvement and Accountability System.

Percentile—the rank in a normal distribution of measurements from 1-100. A score at the 50th percentile is in the middle of the distribution, at the 67th percentile in the upper third.

Program performance—the score a preparation program receives on one or more of the measures used in the Illinois program improvement and accountability system.

Reliable—as a concept in statistics, refers to a source of measurement, an assessment, or other data source that yields similar results under similar circumstances

Sample attrition—loss of members of a group being studied or assessed as representative of a larger population. Significant attrition from the sample compromises the validity of conclusions about the larger population that are based on the sample studied.

School and classroom effects—specific characteristics of schools and their students (e.g., level of poverty, teacher turnover) that can impact the performance of students and their teachers and that need to be taken into consideration when comparing student or teacher performance.

Standardized—also called "normalized", this refers to process of converting all scores on an assessment into scores on a scale from 1-100. This allows scores on different assessments to be compared to one another.

Statistical model—mathematical formula that proposes to explain the complex relationship between a number of variables that can impact an outcome or score. The calculation of teachers' impact on students, for example, often employs a statistical model.

Statistical power—refers to the degree to which the result or outcome for a sample population in an empirical study (e.g., that young children responded better to female teachers) can be confidently judged to be more than a chance correlation in that study only. In general, the larger the sample, the larger the statistical power.

Student growth model—statistical model that predicts an individual student's expected learning gains in an academic year and, when used to evaluate a teacher's impact, considers the difference between the predicted learning gains and the actual gains for all a teacher's students.

TBD—acronym for To Be Determined.

Triangulation—using data from multiple sources (e.g., assessments) to make more confident judgments about an outcome. In evaluating a teacher's strengths and weaknesses, for example, we might draw on supervisor assessments, students' assessments, classroom observations, and student achievement scores.

Valid—as a concept in statistics, refers to the correctness of a source of measurement, an assessment, or some other data source.

Value-added model—refers to a specific kind of statistical model that seeks to determine a teacher's unique contribution to their students' learning in distinction to the impact of other school factors or factors like family or socioeconomic status. The model compares the learning gains of each teacher's students to the average gains of all their peers.