Design of an Education Professional Standards Board (EPSB) Preparation and Accountability System for Teacher Training Programs

Terry Hibpshman
Martin School

March 2013

Introduction

Almost since the beginning of the public school movement in the 19th century, dissatisfaction with the organization and curriculum of the common schools has been a persistent theme in American education.[1] This has led to wave after wave of overlapping reforms.

Until the 1980's, reforms were concerned principally with assuring that the proper *inputs* were available in American schools: physical facilities, curriculum, adequately trained teachers, textbooks, etc. It was implicitly presumed that the various components of the input model were adequate for the purpose of providing quality education, and the principal question was whether schools in particular places had the appropriate resources, or enough of them.

In the 1980's, as a consequence of two reports on the status of American education, there was a qualitative change in the modal approach to evaluation of educational performance. The first report, by The National Commission on Excellence in Education (1983), proclaimed that American education was mired in mediocrity, placing the nation at risk of economic irrelevance. The second, by the Carnegie Forum on Education and the Economy (1986) was critical of the way in which teachers employed by the schools were trained.

Both reports were controversial[2], and numerous authors expressed doubt whether their broad claims about the poor quality of American public education were accurate, but the two reports together set off a new wave of reform that was qualitatively different from previous waves. Since the publication of the two reports, evaluation of American educational systems has focused much more on *outputs* than had been true in the past. The operative question changed from whether the public schools had sufficient resources to do the job, to whether the product being produced was adequate to serve the presumed goal of education as an economic engine.

Since the 1980's there has been increased interest in developing systems that measure the outputs of educational systems, and this trend has accelerated. These measurements are generally labeled "accountability" systems, and they have spread across the country. With the reauthorization of the Elementary and Secondary Education Act in 2002 – the "No Child Left Behind" Act (NCLB) – accountability systems have become mandatory for all states. Even before NCLB, both federal and state efforts to manage the quality of the teacher workforce were already in place. These efforts were principally governed by the 1998 and later amendments to Title II of the Higher Education Act (HEA) (Earley, 2001).

Accountability regimes have since their development in the 1980's gone through a series of transformations as states developed systems that were criticized on technical grounds and replaced by systems that are probably better founded. Yet it is doubtful that anyone has developed an accountability system that is entirely satisfactory on either conceptual or technical grounds. This is true because the great complexity of the educational enterprise, together with limitations of the available measures of both inputs and outputs, always inserts ambiguities into the interpretation of results.

---

[1] For an excellent summary of the history of reform in American education, see Tyack and Cuban (1995), *Tinkering Toward Utopia*.

[2] See for example the criticism by Dorn (1998) of the widely-held perception that the performance of American education is in decline.

There have nonetheless been substantial improvements in both the theory and practice of accountability in education over the past three decades. While there continues to be ambiguity in the results of all systems currently in operation, some principles of adequacy have been established in recent years, and modern systems arguably do a better job of identifying performance deficiencies than did the first systems created in the late 1980's and early 1990's. Application of these principles will not create an accountability model without problems, but will certainly produce a system whose results are more useful than had been true before now.

This report is an effort to apply what is known about accountability system development, based on the experience of state and federal efforts to measure educational system performance over the past three decades, to the performance of the teacher preparation programs that provide educators to Kentucky schools. There are 29 such programs in the state, differing substantially in size and complexity.[3] Any accountability system we might design must take into account these differences in size and complexity, and must also take into account the highly segmented nature of teacher labor markets in Kentucky. It must also respond to the wide differences in educational attainment and community demographics that characterize the various regions of the state.

The report is organized as follows: in part one, we consider what is known about the design of accountability systems, both in education and in other industries. In part two we consider how these principles of accountability system design should influence development of an EPSB accountability[4] system. In part three we consider various measures that might be used to evaluate the effectiveness of teacher preparation programs, together with some ideas about how inferences might be drawn from those measures.

Part one: principles of accountability system design

As used in the evaluation of public services, "Accountability" is not a well-formed concept. The term is used by different authors in different ways, and there seems to be no widely-accepted definition of what is meant by it (Bovens, 2010; Edwards, 2011; Ahearn, 2000). There does seem to be some convergence, however, on a couple of common themes. First, as Romzek and Dubnick (1987) suggest, accountability is a mechanism whereby public agencies manage the expectations of their constituencies (e.g. the public, the legislature, etc.) for services provided by themselves and the service providers they supervise. The second theme defines the relationship between those who provide services ("agents") and those who have an interest in the quality of services provided ("principals") (Edwards, 2011; Dorn, 1998; Bruns, Filmer & Patrinos, 2011; Ewell, 2009; Figlio & Loeb, 2011; Kane & Staiger, 2002; Kim, 2004). The relationship between principal and agent is not always simple or hierarchical: sometimes multiple principals may have an interest in the quality of services provided by some entity (Edwards, 2011), and principals may be either concentrated, as in agencies, or diffuse, as in interest groups defined by status

[3] That is, there are 29 state-approved programs in Kentucky. Historically, about 20% of Kentucky teachers were trained in institutions out of state, which are not the focus of this paper.

[4] EPSB, the Kentucky Education Professional Standards Board, is the state agency with regulatory authority for teacher preparation, teacher certification, and teacher discipline. An accountability system designed to monitor the effectiveness of these three statutory responsibilities is essential to adequate performance of its function.

or demographics, such as the parents of school children (Kim, 2004). The relationships among principals, and among principals and agents, determine how accountability systems are designed.

Various types of principal-agent relationships have been proposed by different authors. Of most interest to us is the idea of "horizontal" accountability (Edwards, 2011), wherein multiple principals share responsibility for the quality of services provided by one or more agents, or more than one level of government is responsible for overseeing the performance of some agent. These are of interest to us because we, together with other state education agencies (in particular The Kentucky Department of Education) share responsibility for the quality of services provided by teachers, and because some of the requirements for educator quality that we are responsible for assuring are imposed by federal legislation.

Authors in this area also define different approaches to accountability system development, depending on who develops and administers the accountability system. Some authors define *professional accountability*, an approach that relies on joint efforts by practitioners to assure the quality of services provided by those in their occupation (Romzek & Dubnick, 1987; Dorn, 2004; Dunn, 2003; Elmore, 2003; O'Day, 2002; Mulgan, 2000). Others define *internal* accountability, systems developed by agents to evaluate the quality of their services independently of accountability systems prescribed by their principals (Brookhart, 2009; Dunn, 2003; Gunzenhauser & Hyde, 2007; Mulgan, 2000; Newmann, King & Rigdon, 1997). Both Mulgan and Newmann, and King & Rigdon suggest that external forms of accountability are likely to fail in the absence of effective internal accountability mechanisms. *External* accountability is defined by these authors as the type of accountability mechanisms imposed by government agencies on schools, districts, or institutions of higher education.

The basic elements required for effective accountability systems vary depending on the author, but there is considerable commonality. Most authors identify the basic elements of an accountability system as including at least some set of performance goals, together with measurements for evaluating whether the goals have been accomplished, and some set of rewards and sanctions for performance (Baker, 2005; Newmann, King & Rigdon, 1997; O'Day, 2002).

An important consideration in goal-setting is whether the goals established by the accountability system are attainable (Abedi, 2007; Briggs & Weeks, 2009; American Evaluation Association, 2006; Ladd, 2001; Linn, 2003). One of the most cogent criticisms of the NCLB AYP goals, as well as the goals of some state accountability systems, has been that the goal of proficiency for all children is simply unattainable, setting a standard that can never be accomplished (Abedi et al., 2007). It has been argued that the effect of setting unattainable goals by NCLB has been to cause states to set the bar for proficiency much lower than is reasonable, in order to avoid sanctions (Armour-Garb, 2008; Manna, 2010).[5]

In principle, goals set by an accountability system should be those outcomes most valued by the principal (Biesta, 2008; Hanushek & Raymond, 2003). In practice, however, some goals may be difficult to measure (Figlio & Loeb, 2011; Dorn, 1998), and as a result we often end up valuing what we can measure, rather than finding ways to measure what we value (Biesta, 2008; Amrein-Beardsley & Barnett 2012; Mangiante, 2011). This defect in accountability design is important because there is substantial evidence that schools and districts respond to the goals set by accountability systems and the measures

[5] So that the goal is actually counterproductive.

used to assess their attainment.[6] If we create accountability systems that only apply consequences for what we can measure, we run the risk of distorting the allocation of resources, reducing performance on some important goals (Pellegrino, 2010).

A very common criticism of educational accountability systems is that the heavy emphasis on the use of achievement tests, especially in mathematics and reading, causes schools and districts to place undue emphasis on those subjects that are tested, to the detriment of untested subjects and non-academic goals of education (Dorn, 1998; Biesta, 2008; Figlio & Loeb, 2011). In addition, the incentives built into accountability systems cause schools and districts to engage in behaviors that reduce the validity of accountability scores. This phenomenon occurs not just in education, but in accountability systems in many fields.[7] Among the distortions caused by accountability incentives in education are the phenomena of "bubble kids", where schools provide inordinate amounts of services to children most likely to move from one accountability category to another (Rothstein, 2008); "teaching to the test", where schools place over-emphasis on academic content likely to be included in the state assessments (Figlio & Loeb, 2011); school and district emphasis on short-term strategies that are not conducive to long-term improvement (Elmore, 2003); reduction in the breadth of school offerings (O'Day, 2002); and reclassification of students to artificially boost scores (Figlio & Loeb, 2011).

In addition to the distortions caused by poor goal selection or over-reliance on achievement testing, accountability systems can cause problems due to the nature of the measures used. No test can comprehensively represent all of the content of interest in any subject, and test developers must always make decisions about how and what to measure within the limited time available (Abedi et al, 2007; Ballou, 2004). In addition, scaling problems make interpretation of achievement test results difficult (Briggs & Weeks, 2009; Jackson and Page, 2011; Seltzer, Choi, and Thum, 2003; Ballou, 2004). The recommended solution for these problems is to assure that tests and curriculum are closely aligned (Board of Testing and Assessment of the National Research Council, 2009; Fuhrman, 2003; Loeb & Strunk, 2007; National Academy of Education, 2009; Supovitz, 2010; Koretz, 2006; Polikoff, 2010; Amrein & Berliner, 2002; Goe, Bell & Little, 2008). Koretz (2006, 2010) points out, however, that alignment may not solve all of the distortion problems caused by accountability measurement.

The problems of accountability measurement are not unique to education. McClellan & Staiger (1999) and Scholle, Sampsel & Davis (2009) found similar problems in medical accountability systems.

The sanctions and rewards provided by accountability systems can generally be of two types, low and high stakes. One type of low-stakes sanction used both by education and non-education systems is the public display of information, as when the federal government requires nutritional labeling on food items, or states produce "report cards" for schools and districts. High stakes sanctions include such things as school closing, school reorganization, dismissal of individual teachers, and graduation examinations.

Despite the difficulties inherent in aligning goals and measurements, there is substantial evidence that accountability systems result in improvement in services, both in education and elsewhere. Simple informational systems alone have been shown to bring about improvement.

[6] By extension, preparation programs will also respond to what is measured.

[7] See Koretz (2010) for a detailed description of the sometimes perverse results of accountability efforts caused by "Campbell's Law". See also Amrein-Beardsley & Collins (2012) for a specific, egregious example of negative consequences resulting from a poorly designed teacher evaluation system.

Williams et al. (2005) demonstrated improvements in hospital performance as a result of collection of informational measures. Jin & Leslie (2003) demonstrated positive effects on restaurant quality as the result of publication of report cards. Mathios (2000) demonstrated changes in consumer behavior as a result of implementation of product labeling in the salad dressing market. It is likely that publication of school and district report cards have had similar effects, although because these reporting mechanisms were generally implemented at the same time as high-states systems, it is difficult to disentangle their effects from higher-stakes methods. Hanushek & Raymond (2003) were able to show an improvement due to reporting that was somewhat lower than improvements due to high-stakes rewards and sanctions. Bruns, Filmer & Patrinos (2011) found improvements in education system performance in Liberia, Pakistan, and Uganda due to implementation of report cards, and Andrabi, Das & Khwaja (2009) found improvements in Pakistan. Weimer (2001) and Figlio & Lucas (2004) found that school report cards had an effect on housing values in the United States.

Numerous studies have demonstrated that state high-stakes systems do result in improvement in education system performance, although the improvements as reported by the states are probably inflated (Brookhart, 2009; National Academy of Education, 2009; Rockoff & Turner, 2010; Dee & Jacob, 2009; Wong, Cook & Steiner, 2011; Hanushek & Raymond, 2003; Neal & Schanzenbach, 2010; Carnoy & Loeb, 2002). That these are real improvements not due entirely to test score inflation is demonstrated by the fact that most of these studies used NAEP data, for which there are no rewards or sanctions, to evaluate how much improvement had occurred. It is important to consider, however, as Swanson & Stevenson (2002) note, that most of the improvement brought about by state accountability systems is of rather modest magnitude.

Historically, educational accountability results generated by state or NCLB systems have been of two types, "status" measures, which compare the magnitude of an aggregated summary of student performance on some measure from one year to the next, and "growth" measures, which use the difference in individual student scores over one or more years as a measure of school or district performance. The original Kentucky accountability system developed in the early 1990's was of the status type, as is Annual Yearly Progress (AYP) required under NCLB. Status approaches have many problems. The most serious difficulty is that they cannot account for demographic and learning trajectory differences between school populations (Fuhrman, 2003). The composition of student populations with regard to ethnicity, previous achievement, and other education-relevant factors varies widely between schools, and within schools from year to year. As a result, it is never clear that the population of students on which a measure is reported is similar across schools or from year to year within the same school. On the other hand, status measures have the advantage of holding all schools accountable for the same level of performance, an important goal from a political perspective (Figlio & Loeb, 2011).

The type of growth measures that have received the most attention in recent years is so-called "value-added measures" (VAM's). These are measures of the improvement in student test scores from one test administration to the next, based on very complex statistical procedures that may or may not adjust for student characteristics such as ethnicity and socioeconomic status, but always attempt to adjust for students' past learning histories. The rationale for this type of measure is that an understanding of a school's (or teacher's) performance requires knowledge of the circumstances of practice where the learning occurred. VAM's are very controversial for a number of reasons. Because

the data on which they are based are very "noisy", and because we are never sure that we have data on all the factors that might influence student learning, there is considerable uncertainty in the results. Generally, these measures are able to produce reliable results only at the extremes of the distribution of performance (Amrein-Beardsley, 2008). Still, they are arguably better measures than we have had in the past (Goldhaber & Evans, 2007). They have proven very useful as research tools for teasing out the contribution of various factors to educational attainment.

An additional problem with VAM's is that they implicitly set different goals for different teachers and schools (Ballou, 2004). This is a problem because it in effect gives poorly-performing districts, schools, or teachers an excuse for poor performance, and perhaps gives them less incentive to improve.

A popular recent addition to the available growth measures is the "Student Growth Percentile" (SGP), now in use in numerous states, including Kentucky (Buckley & Marion, 2011; Betebenner, 2011; Kentucky Department of Education, 2012). This is a measure of a student's growth in achievement from one year to the next, compared to other students with similar prior achievement, unadjusted for student characteristics. These measures have the advantage of intuitive simplicity, but do not allow for comparisons among teachers.

There are a few additional considerations. First, because sampling errors are larger for small samples than for large ones, small schools and districts are more likely to be perceived, based on accountability measures, as performing either exceptionally well or exceptionally poorly than are large schools and districts (Figlio & Loeb, 2011; Fuhrman, 2003; Hollingshead & Childs, 2011).[8] This is also true of measures of teacher performance: small classrooms will generally have greater variation in accountability scores than will larger ones.

An important consideration in the selection or development of measures of accountability is their sensitivity to changes in performance at the classroom or school level (Kane & Staiger, 2002; Polikoff 2010; D'Agostino, Welsh, & Corson, 2007; Goe, Bell & Little, 2008). This is a feature of accountability measures described by some authors as "instructional validity". It is important because of a problem noted by some authors: in order to respond to an accountability measure, an agent must have the ability to identify what practices are associated with performance (Nichols, Meyers & Burling, 2009). This capacity will be enhanced if the agent has an effective internal accountability system.

A related consideration is that of organizational capacity. Educational accountability systems make the implicit assumption that failure to meet goals is a consequence of inadequate effort or inappropriate allocation of resources (Koretz, 2006). This is sometimes the case, but in some cases the problem may be that the organization lacks the capacity, either though limited resources or deficient knowledge, to make the improvements necessary to avoid sanctions (Brookhart, 2009; Ahearn, 2000; Biesta, 2008; Elmore, 2003; Figlio & Loeb, 2011; Loeb & Strunk, 2007; National Academy of Education, 2009; Gunzenhauser & Hyde, 2007; Ladd, 2001; Newmann, King & Rigdon, 1997). This is a general principle in accountability systems: an agent should be accountable only for those things it is able to control.

Accountability systems should themselves be subject to ongoing evaluation (Baker, 2005; Ahearn, 2000; Supovitz, 2010; Amrein-Beardsley & Barnett, 2012; Ananda & Rabinowitz, 2003). Given the numerous limitations and difficulties noted above, it is certain that any accountability system will

[8] That is, measures for small units are relatively unstable.

have problems and will be likely to generate unintended consequences. Development of these systems requires an ongoing effort to assure that measures are appropriate, that unintended consequences are limited, that goals are appropriate to the mission of the system, and that new goals and measurements are recognized and incorporated into the system.

A final promising idea comes originally from the airline industry, the idea of a "just culture". People who work in industries, such as the airline industry and the health care industry, where poor performance can have catastrophic consequences, have developed an approach to accountability designed to assure that errors are reported and addressed, despite the very understandable tendency people have to avoid sanctions (Frankel, Leonard, & Denham, 2006; Furger, 1997; Gain Working Group, 2004, Ewell, 2009; O'Day. 2002). Just culture eschews both "cultures of blame", where every error is sanctioned, and "blameless cultures", where no error is sanctioned. Instead, just culture divides types of errors into those that are merely errors in judgment or errors due to system malfunction from those due to negligence, and of most importance, to recklessness. It prescribes a range of remedies, and sanctions only egregious conduct. It gives precedence to the expertise of those with the greatest knowledge relevant to a problem, rather than to hierarchical structures of authority. The idea is to remove the fear of severe sanctions for honest errors, while discouraging irresponsible conduct. It is believed that the current very low rate of airline accidents is traceable to the implementation of this idea in the airline industry.

Part Two: Considerations necessary to EPSB accountability design

A number of accountability issues specific to teacher training and certification need to be discussed before an EPSB accountability methodology can be developed. Some accountability efforts have already been attempted at both the state and federal levels, beginning around 2000. These efforts have not been very successful, but they have provided experience that should help guide us as we develop a new approach.

Since the 1990's, attention has increasingly shifted to evaluation of the performance of teachers, following the claim by Sanders and his collaborators (Sanders & Horn, 1994) that the single largest factor in student achievement is teacher quality. This sweeping statement has been echoed by many, but later authors have modified it to state that teacher performance is the greatest education factor amenable to administrative management (Hibpshman, 2012; Goldhaber & Hansen, 2012). Whether one believes Sanders' sweeping portrayal of the importance of teacher performance, or the more cautious assessment of later authors, both federal and state education agencies have increasingly since the 1990's attempted to manage the quality of the teacher workforce.

Evaluation of the performance of individual teachers is one mechanism for evaluating the effectiveness of preparation programs. Whether it is true that preparation programs serve a gatekeeping function only, or whether they "add value" by developing candidate skills beyond what would be predictable by talent alone, is of little interest to district personnel managers, who are only concerned with whether they are hiring effective staff (Raudenbush & Willms 1995), but the difference is of great importance to us. We are of course concerned about the capacity of preparation programs to select suitable candidates, but if preparation programs add nothing beyond that, there is no real reason to have preparation programs at all. This question was the impetus behind Noell's work in Louisiana

(Noell, 2007). He was able to demonstrate differences among programs, but many authors in the teacher effectiveness literature have found little reason to believe that teacher training matters much (Mason, 2010; Goldhaber & Liddle, 2012; Osborne et al., 2013). A few authors (Clotfelter, Ladd, & Vigdor 2007a, 2007b) have found substantial effects for teacher preparation.[9]

Federal attempts to manage the quality of teachers were first authorized in the amendments to Title II of the Higher Education Act (HEA) in 1998 (Earley, 1998). These amendments provided for a variety of measures intended to improve teacher quality nationwide, but the most important were reporting requirements by both states and teacher training programs. Title II requirements became substantially more complex and demanding after 2009 (USDOE 2013).[10] Subsequently, the No Child Left Behind Act required reporting by states of the "highly qualified" status of teachers employed in schools that are funded in part via Title I of the Elementary and Secondary Education Act. HEA Title II established a reporting system for teacher training programs that displays process measures thought to be associated with teacher quality, such as the proportion of teacher candidates who passed the relevant state teacher tests (in the case of Kentucky, the Praxis II tests). NCLB established a very complex set of criteria for determining when a teacher is highly qualified. These criteria depended heavily, as do most NCLB requirements, on state definitions, but generally reflected whether a teacher had been adequately prepared in the content area they were assigned to teach.

Neither the Title II nor the NCLB accountability mechanisms were very successful in improving the quality of the teacher workforce. In the case of Title II, the principal measure of preparation program quality – passing the Praxis tests – was not well defined, and was easily manipulated by preparation programs. As a result, it was unusual for any institution to have a proportion passing that fell below the cutoff, and many institutions regularly reported 100% pass rates for all relevant Praxis tests (Hibpshman, 2006).[11] The NCLB requirements had similarly vague criteria, which allowed states to define almost all teachers as highly qualified, even though the population of candidates changed very little (Columbia University Teacher's College, 2009).[12]

EPSB responded to both initiatives with an elaborate system for collecting the data required to comply with federal requirements. In addition, EPSB created its own accountability system, the Quality Performance Indicator (QPI), an index that combined information about Praxis pass rates, internship completion rates, and ratings on the new teacher survey that was implemented in 2003. 60% of the

[9] This illustrates the specification problem in studies of teacher effectiveness. Because of the great complexity of the problem and limitations in the available data, different models give different results, and it is not always clear which is a better model.

[10] The increase in the complexity of the Title II requirements is best illustrated by the number of different titles required in the state report. From 2002-2008 there were 8 titles; in 2009 there were 7; in 2010 there were 25; and in 2011 there were 26.

[11] If preparation programs do not vary in their apparent performance, low-stakes accountability systems such as the Title II report card are unlikely to have any effect, since there is never any reason to believe that a program is performing poorly. That is, there is no "embarrassment factor" to motivate programs to make changes.

[12] For example, in Kentucky, many of the people who would previously have been issued emergency certificates – a practice strongly criticized by the federal Department of Education – instead entered alternative preparation programs, which were acceptable under the HOUSSE provisions of the "highly qualified" rules. Yet they were the same population of people, and entered the classroom with no more delay than had they been issued emergency certificates.

index was based on Praxis pass rates, and 20% was based on each of the other two measures.[13] This index proved unsatisfactory for a number of reasons. Most importantly, it suffered from the same problem of inflated levels of Praxis pass rates as had the Title II system. Since more than half of the index was based on Praxis pass rates, most institutions got about half of the available points. Additionally, because internship failure rates are remarkably low (generally less than 1%), most institutions got all of the available points for internship pass rates. The new teacher survey results seemed subject to a halo effect, which gave most institutions high marks on that indicator as well. The QPI thus served as an insensitive measure that rarely showed any institution to be performing poorly, even when agency staff were aware of severe problems (Hibpshman, 2006). The QPI was abandoned in 2009 as result of these problems. An additional problem with the QPI was that, as with all statistical measures, it produced unstable measures for small programs, and could not be used at all as a measure of programs with very small numbers of graduates.

Since 2009, a number of ideas about development of a new accountability system to replace the QPI, and to supplement the insensitive federal approach to accountability, have been under consideration. These efforts have amounted to proposals for development of additional measures of teacher training program performance, to be incorporated into a "dashboard" reporting system, as well as a high-stakes accountability regime (Hibpshman, 2010). Some of the ideas for the dashboard have been under development since 2011. The high stakes system has yet to be developed or implemented.

Since 2011, the Kentucky Department of Education has been engaged in the development of a "Professional Growth and Effectiveness System", incorporating multiple measures of teacher effectiveness (Kentucky Department of Education, 2013). This system is currently under development, and will not be fully implemented before the 2014-2015 school year. It includes a "growth" measure based on SGP's, a standard teacher evaluation, student ratings of teacher effectiveness, teacher reflections on their own performance, and other measures. It is unclear at this point how these various measures will be used to evaluate teacher effectiveness, or whether they will be of sufficient reliability and validity to be used in a high-stakes environment.

Accountability measures developed by the federal government, and by KDE, are important to us because we exist in an environment of shared accountability for teacher performance (a "horizontal" accountability environment, as described above in section one). Some requirements, such as those of NCLB and Title II of HEA are imposed on us by the federal government, and we are obliged to comply. The KDE teacher accountability system is important to us for two reasons. First, some of the data necessary to our evaluation of teacher preparation programs must necessarily come from KDE, either from their teacher accountability system or from other relevant KDE data stores;[14] secondly, it will be necessary as we develop our own accountability system to define the division of responsibility between ourselves and KDE. The latter consideration is important: with the exception of teacher discipline, we have no responsibility for the performance of individual teachers, except insofar as we are responsible for assuring that teachers supplied to schools by EPSB-approved preparation programs are of sufficient

[13] Indexes suffer from what is known in the economics literature as the "Index problem", the difficulty in accurately setting weights.

[14] This may well be indirectly, from the P-20 data collaborative.

quality. Yet surely some of our evaluation of preparation program performance will rely on measures of the performance of individual teachers.

One consideration we must take into account is the question of attribution for teacher performance. Newly-hired teachers are likely to be heavily influenced by their preservice training, but as they progress through their careers, other factors, such as professional development and the day-to-day practice of teaching, become more important. Studies of the effect of teacher preservice training seem to indicate that institutional effects are not likely to be a major factor in teacher performance after the first few years (Noell, 2007). For this reason, we should be reluctant to hold preparation programs accountable for teacher effects beyond the first few years following training, and should weight earlier years post-completion more heavily than later years (Walters-Parker, 2012). Incorporation of some sort of proportional accountability scale, indexed by the number of years post-completion, would be a highly attractive element of any system we might develop.

In addition to federal and other state agency imperatives that govern how we should implement an accountability system, we need to consider the recommendations of national organizations with an interest in preparation program accountability. In particular, we should consider the recommendations of the Council for the Accreditation of Teacher Preparation (CAEP) (2013).[15] These recommendations were recently released, and contain a great deal of important information. The standards include standards for content and pedagogical knowledge; clinical partnerships and practice; candidate quality, recruitment and selectivity; program impact; and provider quality, continuous improvement, and capacity. In addition to these proposed standards, the draft document recommends a data acquisition and reporting mechanism by CAEP to monitor the quality of teacher education programs nationally.

The Obama administration, dissatisfied with the use of input measures as indicators of teacher preparation program quality, recommended in 2011 a set of output measures based in part on student growth (Coggshall, Bivona & Reschly, 2012). These measures included aggregated measures of the learning outcomes of students taught by the graduates of teacher preparation programs, assessment of job placement and retention rates of graduates of teacher preparation programs with attention to shortage areas, and collection of surveys of program graduates of their perception of performance and effectiveness of the programs from which they graduated.

An important consideration in evaluation of preparation programs is that the teacher labor market in Kentucky Is heavily segmented. Our previous evaluation of this issue has demonstrated that about three-fourths of the districts in Kentucky are “dominated” by a single teacher training program. That is, a large proportion – often half or more – of the teachers employed by the district were trained by one institution (Hibpshman, 2004). This is important because it places severe constraints on our ability to evaluate the differential effectiveness of teachers trained by different programs. In a study of teacher effectiveness in Kentucky in 2011 (Kukla-Acevedo, Streams & Toma, 2011), researchers at the University of Kentucky concluded that the heavy segmentation of the teacher labor market precluded use of many of the more popular models for evaluating program performance. This problem was also noted in Louisiana (Noell & Burns, 2006) and elsewhere (Floden, 2012). The problem is important because it means that we will have to use very complex and sophisticated methods to sort institutional

[15] CAEP was formed by merger of the National Commission on the Accreditation of Teacher Education (NCATE) with the Teacher Education Accreditation Council (TEAC).

programs. One model that has been suggested is *propensity scoring*, a method that compares institutional programs by finding teachers from different programs who are similar except in their association with a particular institution. This is a valid approach to the problem, but considerable expertise is required (Abadie & Imbens, 2011; Coca-Perraillon, 2006; Heinrich, Maffioli, & Vázquez, 2010; Kelcey, 2011; Parsons, 2004).

As the CAEP standards and the proposed federal standards note, in addition to consideration of issues related to the quality of individual teachers, we have an interest in a number of matters related to preparation program quality. Regardless of whether teachers as individuals trained by approved programs are of good quality, we should be concerned about the following:

- Whether enough teachers are being produced statewide to avoid shortages in particular content areas
- Whether preparation programs are doing an adequate job of screening applicants for admission
- Whether preparation programs are aware of, and adequately respond to, specific conditions in districts where they have a strong presence
- Whether preparation programs are responsive to the need for teachers in particular content areas, especially those that are perceived to be in shortage
- Whether preparation programs meet the requirements of accreditation agencies
- Whether preparation programs are well-managed
- The proportion of candidates who actually apply for, and are granted, certification
- The proportion of candidates who are employed as teachers within some reasonable time following program completion
- The proportion of alternative program candidates who complete their program
- Whether teachers, once they are in the classroom, persist in education
- Whether additional training provided by preparation programs to teachers after certification is relevant and of adequate quality
- Process measures, such as the average length of time from admission to completion, the mix of content delivered by programs, the proportion of admitted candidates who complete within a reasonable time, the proportion of candidates who are full time students, etc.
- The existence and quality of internal accountability systems within teacher preparation programs

The above, together with the question of whether programs are producing effective teachers, argues for a three-tier model of accountability, including:

- Measures of program management
- Measures of program processes
- Measures of the effect of programs on local educational systems, either at the district/school level, or through the performance of individual teachers.

Global design considerations

The following seem reasonable as general principles for design of an accountability system:

- Any system we design must take into account the horizontal nature of accountability for teacher effectiveness
- Measures must be selected with an awareness of particular conditions – such as the highly segmented teacher labor market – that might reduce or obviate their validity
- Measures of accountability should address those things within the control of preparation programs, and should be able to differentiate between circumstances where poor performance is a consequence of bad program design or poor judgement, and those where the program lacks the capacity to respond to performance deficits
- To the extent possible, accountability measures should be practice-relevant. That is, a program should be able, on receiving results from the system, to identify those practices that are most likely to have resulted in poor performance, or to identify practices that are likely to result in better performance
- Development of an EPSB accountability system should be accompanied by incentives and technical assistance to help programs develop effective internal accountability systems
- To the extent possible, sanctions for poor performance should be administered within a "just culture" framework
- Any accountability system we might develop should be subject to ongoing review and revision

Part Three: Possible measures and methods

We begin this section with a discussion of existing EPSB goals, as published on the EPSB website (Kentucky Education Professional Standards Board, 2012a). The goals aim at the achievement of a vision and mission statement:

> Vision Statement
> Every public school teacher and administrator in Kentucky is an accomplished professional committed to helping all children become productive members of a global society.
>
> Mission Statement
> The Education Professional Standards Board, in full collaboration and cooperation with its education partners, promotes high levels of student achievement by establishing and enforcing rigorous professional standards for preparation, certification, and responsible and ethical behavior of all professional educators in Kentucky.

In pursuit of this vision and mission statement, EPSB has established five overarching goals, each of which has a number of strategies:

Goal 1: Every approved educator preparation program meets or exceeds all accreditation standards and prepares knowledgeable, capable teachers and administrators who demonstrate effectiveness in helping all students reach educational achievement.

Strategy 1.1. Maintain regular and rigorous reviews of all program quality indicators.
Strategy 1.2. Document and publish information on the quality of each preparation program.
Strategy 1.3. Provide technical assistance to support program improvement.
Strategy 1.4. Utilize research to inform program improvements.
Strategy 1.5. Review programs to ensure focus on student learning.
Strategy 1.6. Maintain a focus on continuous improvement of all preparation programs.
Strategy 1.7. Provide accurate and reliable data to support decision making.

Goal 2: Every professional position in a Kentucky public school is staffed by a properly credentialed educator.

Strategy 2.1. Document every assignment of educators in Kentucky public schools.
Strategy 2.2. Document the highly qualified status of all Kentucky teachers as required under NCLB.
Strategy 2.3. Monitor the validity and reliability of teacher and administrator assessments.
Strategy 2.4. Document and publish the results of all assessments required of new teachers and new administrators.
Strategy 2.5. Maintain a focus on continuous improvement of all traditional and alternative route certification procedures and processes.
Strategy 2.6. Provide accurate and reliable data to support decision making.

Goal 3: Every credentialed educator exemplifies behaviors that maintain the dignity and integrity of the profession by adhering to established law and EPSB Code of Ethics.

Strategy 3.1. Promote awareness of the EPSB Code of Ethics.
Strategy 3.2. Maintain an accurate database of misconduct and character and fitness cases.
Strategy 3.3. Present in a timely manner all cases for review by the EPSB.
Strategy 3.4. Maintain a focus on continuous improvement of all hearing procedures.
Strategy 3.5. Provide accurate and reliable data to support decision making.

Goal 4: Every credentialed educator participates in a high quality induction into the profession and approved educational advancement programs that support effectiveness in helping all students achieve.

Strategy 4.1. Develop and utilize reliable measures of teacher effectiveness and student achievement that may be used in evaluation of induction and professional advancement activities.

Strategy 4.2. Ensure that every new teacher and principal has a high quality induction experience while demonstrating knowledge and skills that support student learning.

Strategy 4.3. Ensure that high quality mentoring and support services are provided for teachers seeking National Board for Professional Teaching Standards certification.

Strategy 4.4. Ensure that the Continuing Education Option for rank change program maintains appropriate rigor while demonstrating advanced knowledge and skills that support student learning.

Strategy 4.5. Provide accurate and reliable data to support decision making.

Goal 5: The EPSB shall be managed for both effectiveness and efficiency, fully complying with all statutes, regulations and established federal, state, and agency policies.

Strategy 5.1. Maintain a qualified and diverse EPSB workforce.

Strategy 5.2. Ensure that all personnel are experiencing life-long learning and professional experiences that support their professional growth.

Strategy 5.3. Seek full funding for all EPSB operations, personnel, and programs through an approved biennial budget request.

Strategy 5.4. Provide semiannual budget reports to the EPSB.

Strategy 5.5. Maintain facilities, equipment, and agency technology that support efficient and productive agency operations.

In addition to the above, the EPSB Program Guidelines document (Kentucky Education Professional Standards Board, 2012b) describes “themes” that preparation programs are expected to pursue in the process of program development:

- Diversity (with specific attention to exceptional children including the gifted and talented, cultural and ethnic diversity)
- Assessment (developing skills to assess student learning)
- Literacy/Reading
- Closing the Achievement Gap (identify what courses emphasize strategies for closing the gap)

An important distinction for our purposes is between goals as general statements of desired outcomes, as opposed to goals as specific target quantities. The goals and themes as described above are consistent with the former: however important they may be, it would be difficult to determine, as written, whether any of them had been met. In order to operationalize these or any other goals we might want to accomplish, it is necessary to establish target values. Such values are always to some extent arbitrary, and are not always easy to establish.

Once target values for our various goals have been established, it is necessary to establish what consequences, if any, should be applicable when a preparation program fails to achieve the targets. In keeping with the discussion of "Just Culture" above, we are inclined to divide deficiencies that might be identified by accountability measures into two general categories, those that result from inadequate resources or errors in judgement, which are likely to be relatively minor, and those that are so serious that sanctions may be required. The type of consequences proposed here are generally consistent with these categories, which are listed here in order of severity:

- "Report Card" consequences involve only reporting of measures of program performance, based on the finding that simple reporting often brings about improvement
- "Technical assistance" involves discussion between EPSB and preparation program staff of the source and possible remedies for a deficiency, and design/implementation of new program elements, or modification of existing elements, intended to bring about remediation
- "Program review" involves EPSB staff investigating a preparation program to determine the source of a deficiency, and development by EPSB staff of a recommended solution
- "Sanctions" would involve one of three increasingly severe consequences for poor performance:
    - Identification of the program as a poorly-performing program
    - Limitations on the program's privileges (e.g., limitations on the level, content areas, or geographic locations permitted to the program)
    - Decertification

Although the above philosophy generally tries to impose consequences at the lowest appropriate level of severity, it is necessary that the system have an escalation component. When a problem with minimal consequences, such as a Praxis II pass rate below the acceptable level, persists despite gentle efforts at remediation, more severe consequences should be contemplated.

Tables 1-3 present the design of the proposed accountability system. Table 1 describes the 25 proposed measures in general terms. Table 2 cross-indexes the proposed measures to the EPSB goals

and strategies. Table 3 defines the measures listed in table 1 in more technical terms, and adds target values where appropriate.[16]

We note here that the number of measures is large, and implementation of any significant number of them will require substantial work. Some of the proposed methods are already in place or under development. We already have a program accreditation component, and a data dashboard that includes some of the measures is under development. The latter will require some additional display elements. Accountability elements that will require new development include principally those that involve complex statistical models, and those proposed by the federal government.

The development of the more complex models presents some problems. No one on staff at EPSB has the expertise to develop or manage these models; the agency must either acquire the staff to develop these methods, or find a vendor to develop them. Data are also a problem. It would be possible at present to develop a value-added model to compare teachers across preparation programs, using propensity scoring as a method, and using either the accountability scaled scores or SGP values as reported by K-PREP. It would also be possible to develop teacher effectiveness measures using other data such as high school graduation or enrollment in advanced courses, using hazards models. Development of measures of teacher performance from the KDE PGES system now in development is more problematic, since PGES results will not be available for all districts before the 2014-2015 school year. Development of disaggregated measures using SGP values, as proposed by the federal government, will require considerable effort. Simple comparison of disaggregated SGP values will not be an unbiased measure of preparation program performance.

A possible strategy for developing complex accountability methods is to develop a set of methods now, using the data available, that could be used to evaluate any of several different types of outcome data. As suitable data become available in the future, these methods could then be applied to them. This would allow us to test different outcome data for suitability on an ongoing basis. The amount of effort required is significant, but it is apparent that we will have to do this kind of analysis eventually.

The general strategy we propose for development of the accountability system is as follows:

- Create scales for the reporting of existing methods, such as accreditation processes
- Identify additional elements for the dashboard, and develop display formats
- Develop methods for creating the disaggregated values using SGP data
- Develop generalized statistical procedures for analysis of longitudinal data
- Develop procedures for review and revision of the accountability model

In addition, it will be necessary for EPSB staff to set target values for each of the accountability measures where they are appropriate, and to refine the consequences model.

[16] The target values in table 3 are my own best guess as to what would be a reasonable value, based on my experience with EPSB data over the years. The actual targets to be used in an accountability system will ultimately be a matter of negotiation among Board members and preparation program staff.

Inferential procedures

Assessing the performance of teacher preparation programs is an inferential process. That is, our accountability system will always make informed decisions based on incomplete and noisy evidence. Recent authors in the area of educational accountability measurement have concluded that inference based on a variety of measures may be an attractive alternative to the use of sanctions based on causal models that depend on a few measures (Bettebenner & Linn, 2009; Pellegrino 2010).

Inference, by its nature, is a process of arriving at a conclusion that goes beyond the available evidence (Krynski & Tenenbaum, 2007). This is an example of inductive reasoning, the process of arriving at a belief about reality based on evaluation of particular facts. Conclusions drawn on the basis of inductive reasoning are always to some extent uncertain (Chater, Oaksford, Hahn & Heit, 2009), they are always to some extent subjective, and they are *nonmonotonic*: conclusions may change based on additional evidence (Kyberg & Teng, 2001).

That our decision-making about program performance is an inductive process subject to the above limitations does not mean that it is arbitrary or capricious. To the extent that we create measures of the goals we hope to achieve, and evaluate the results within a consistent decision-making framework, we are likely to make better decisions about program quality than we would without such a framework. If we monitor the performance of the accountability measures and make adjustments when necessary, we can deal effectively with the problems of uncertainty and nonmonotonicity.

The proposed measures described herein cannot be evaluated with any single methodology. Some are more subjective than others, and some are entirely qualitative. Some of the measures can be evaluated with a "bright line" criterion, where failure to achieve some minimum quantitative measurement is presumptive evidence of poor program performance, but not all can or should be subject to this sort of criterion. Because of the large number of proposed measures and the fact that they cannot all be evaluated within a single inductive framework, we propose a multivariate decision-making strategy that employs multiple methods while avoiding any single measure of program performance. Using this proposed approach, a preparation program might be judged to be ineffective for any of several independent reasons despite several indicators of program strength, or might be judged to be effective overall despite the existence of identified deficiencies.

We propose specifically the yearly production of a *program performance report*, incorporating each of the measures, which would be analyzed by EPSB staff, using a variety of methods as appropriate, to arrive at conclusions about program quality and proposed consequences for identified deficiencies. The methods used to evaluate the various measures would be one of the following, depending on the type of measure at issue:

- Qualitative decisions based on some rubric of adequacy
- Decisions based on quantitative measures where no bright line can be established
- Decisions based on quantitative, bright-line measures
- Informational results where no performance criterion is appropriate or can be reasonably established

Where possible, we propose that quantitative judgements be based on charting methods. Performance charting has a long history in other fields, particularly in the management of industrial processes, and has in recent years been successfully adapted to the evaluation of service organizations, including education (Woodall, Adams & Benneyan, 2011; Schafer, Coverdale, Luxenberg & Jin, 2011). Well-established procedures for making decisions using this methodology exist, and it has been demonstrated that these methods can be used effectively by staff who lack sophisticated statistical training (Omar, 2003). Some modification of traditional procedures will have to be made for our purposes. One of the advantages of this type of analysis is that charting makes it possible to infer a relationship between particular program innovations and outcomes (Woodall, Adams & Benneyan, 2011).

Conclusion

Accountability systems are mechanisms for determining whether goals set by administrative agencies (principals) have been met by organizations (agents) responsible to those agencies. It is not always easy to develop effective measures of some goals set by the principal, but it is important to assure that program distortion does not occur due to emphasis on those goals that can be easily measured. All accountability systems suffer to some extent from problems such as unintended consequences and ambiguity due to measurement error. These problems can be substantially reduced by careful attention to system design, and periodic adjustment of the accountability measures.

Any system we might implement must be developed in a horizontal accountability framework, because of the shared responsibility between EPSB and KDE for teacher quality, and requirements imposed by federal statute. This will require coordination of our efforts with other agencies, and the use of data supplied by other agencies.

EPSB has established five overarching goals over the years, which serve as the basis for the development of an accountability system for preparation program quality. These goals are further articulated through a series of strategies, which might better be thought of as subgoals. Not all of the strategies are germane to preparation program accountability, though most are. In addition to the goals, EPSB defines “themes”, which serve as guidelines for program development. We add here a few additional goals that we believe to be implicit in EPSB’s mission. Together, these strategies, themes, and additional goals constitute a reasonably comprehensive definition of what it means to be a preparation program of sufficient quality. It is acknowledged that not all of the goals identified herein are of equal importance, and we operationalize this principle by specifying different levels of consequences for inadequate performance.

We propose that the accountability system be implemented within a “just culture” framework, where a distinction is made between errors due to inadequate resources or honest mistakes, and those due to negligence or recklessness. Consequences for the former type of error would generally be less severe than for the latter, although we also suggest there be an escalation component of the system, to guard against persistent failures. This escalation component is based on the idea that a persistent failure, in the face of ongoing efforts at remediation by EPSB, is most likely due to negligence.

Implementation of the ideas represented above will require time and substantial effort. Some of the proposed methods will require the involvement of persons with technical expertise beyond that

likely to be present among EPSB staff. It will be necessary as we begin to develop this system to determine, for each of the proposed measures, whether the necessary data exist, whether the measure can be developed at a reasonable cost, the relative importance of the measure, and who should develop it. That is, we should develop a detailed implementation plan.

Regardless of whether we implement the approach contained herein or some other, it is certain that we will need to develop a new accountability system. Whatever system we create will have to do the following:

- It will have to comply with federal requirements under Title II of HEA and NCLB, and perhaps other federal guidelines
- It will have to be coordinated with KDE teacher performance measurement
- It will require much more sophisticated types of analysis than we have used in the past
- It will require multiple measures and complex decision-making

Table 1
Accountability Measures

Item number	Principle	EPSB Goal	Measure	Comments
1	Teacher quality	Goal 2, Strategy 2.2 relates to NCLB highly qualified status. Goal 1, Strategy 1.2 and 1.5 relates to teacher effectiveness	Student achievement as adjusted for student and school characteristics and perhaps other factors. Perhaps also such things as teachers' effect on student graduation (requires a hazards model), or their effect on whether students subsequently take advanced placement or dual credit courses (also requires a hazards model).	SGP's might serve as the outcome variable, or it might be necessary to use K-Prep scaled scores. It will almost certainly be necessary to use propensity scoring to compare across programs. To the extent that data are available, this may be a better way to address the small program problem. Greater weight should be given to teachers in the first few years of practice, and teacher performance beyond a cutoff (probably 5 years) should not be attributed to preparation programs.

2	Teacher performance appraisal	Goal 1, Strategy 1.2 relates to teacher effectiveness	KDE PGES data	Not possible as an operational measure before the 2014-2015 school year. Data to create and test EPSB measures may be available prior to that. Will depend heavily on the reliability of the data produced, which has not yet been established. As with raw teacher performance, should be weighted to account for the number of years post-program completion, and preparation programs should not be held accountable for teachers beyond a reasonable cutoff.
3	Preparation program attention to district/school characteristics	Relates obliquely to the "themes" in the Program Guidelines document, but EPSB has no specific goal in this area. In addition to the specific categories contained in the themes, we will need to establish goals for such things as district/county/regional education levels, etc.	This will require an ongoing assessment of the market share by each program in schools and districts. Programs will have to provide EPSB with information about efforts they make to address the educational problems of schools and districts where they have a strong presence.	Will require some sort of discontinuity design. Some of the evaluation will necessarily be qualitative.
4	Productivity	Relates obliquely to the "themes" in the Program Guidelines document, but EPSB has no specific goal in this area.	The number of teachers produced each year in different content areas, compared to the number needed in the schools and districts where the program has a strong presence.	Development of some sort of projection of the number of teachers needed will have to be developed. Improvements in the admissions and exit system will be needed, to make sure we have an accurate understanding of the status of the pipeline at any moment.

5	Candidate screening	Goal 1, Strategy 1.6, and Goal 2, Strategy 2.5 relate obliquely to this measure	The proportion of accepted candidates who required developmental classes compared to the institutional proportion. Mean candidate GPA. Mean candidate PPST scores. Analysis of "dispositions" screening. Number of candidates denied for various reasons. Mean candidate ACT scores. In the case of graduate students, mean GRE scores.	Data for all of these can come from the P-20 system, but data for non-public institutions may not be available in P-20, and a supplementary system may be required.
6	Accreditation status	Goal 1, Strategy 1.1	Accreditation reviews from CAEP or EPSB	It would probably be helpful to develop some sort of scale for these.
7	Program management	Goal 1, Strategy 1.1	Accreditation reviews from CAEP or EPSB	It would probably be helpful to develop some sort of scale for these.
8	Candidates completing within a criterion length of time	Goal 1, Strategy 1.2	Proportion of admitted candidates who complete	The horizon should probably be either 3 or 4 years. It should be recognized that the target for this measure will be substantially less than 1.
9	Completed candidates who apply for certification within some time horizon	Goal 1, Strategy 1.2	Proportion of completers applying for certification	The horizon should probably be 3 years. Inferences will have to take conditions in specific teacher labor markets into account.
10	Completed candidates who are employed within some time horizon	Goal 1, Strategy 1.2	Proportion of completers employed	The horizon should probably be 3 years. Inferences will have to take conditions in specific teacher labor markets into account.

11	Completed candidate Praxis II pass rate (Global Praxis pass rate)	Goal 1, Strategy 1.2, Goal2, Strategy 2.4	This is the proportion of completers who passed the Praxis II, regardless of the number of administrations	Programs should not be allowed to define admission or completion status by Praxis pass status. A time horizon will have to be set, or some procedure to deal with data censoring will have to be developed.
12	Raw Praxis II pass rate	Goal 1, Strategy 1.2, Goal2, Strategy 2.4	This is the proportion of admitted candidates who passed the Praxis II tests within some criterion time of admission, regardless of whether they were counted as completers	This is a check against the completed candidate pass rate, to guard against gaming the system. Some institutions, for example, are inclined to not count an admitted candidate as a completer until he/she has passed the relevant Praxis test, thus guaranteeing a perfect score on the global Praxis pass rate.
13	Raw Praxis II first-time pass rate	Goal 1, Strategy 1.2, Goal2, Strategy 2.4	The proportion of admitted candidates who passed the relevant Praxis II test on the first administration	This is controversial among preparation program staff, but it makes good sense. Because we are concerned with value added by the institution and not with value added by other entities, and because candidates who fail are likely to receive assistance from other entities, it is a purer measure of institutional effectiveness than the other Praxis II pass rates.
14	Conditional credentials issued	Goal 1, Strategy 1.2, Goal2, Strategy 2.4	Number/proportion of program completers issued conditional certificates	This will vary with the Praxis pass rates

15	Review of assessments	Goal 2, Strategy 2.3	Yearly documentation that a review was undertaken and completed, together with an assessment of the validity/reliability of assessments.	Not a direct measure of preparation program performance, but essential to interpretation of program performance results. This is already in effect, and just needs to be documented as part of the accountability system.
16	Preparation productivity by type of program (e.g., alternative or traditional)	Goal 2, Strategy 2.5	Number and proportion of candidates admitted by content area and program type. Number and proportion of admitted candidates by content area and program type completing within some time horizon. Number and proportion of admitted alternative candidates who subsequently entered a traditional program without first completing the alternative program. Number and proportion of alternative candidates who subsequently transferred to another alternative program.	This will result in some estimates based on very small samples, which will be difficult to interpret.
17	Program management of clinical/induction support	Goal 4, Strategy 4.2	Mean number of support hours per candidate, compared to some criterion of adequacy. Results of audits. Results of complaint resolution.	This will require some additional data collection beyond what is now available.

18	Annual review of the accountability system	Goal 1, Strategies 1.4 and 1.7; Goal 2, Strategy 2.6; Goal 3, Strategy 3.5; Goal 4, Strategies 4.1 and 4.5	Review of technical measures such as standard errors and effect sizes from the different measures used to evaluate program effectiveness. Review of current research to determine whether alternative measures are available. Trial calculation of alternative measures.	This requires analysis by someone with advanced technical skills.
19	Review of preparation program content	Goal 1, Strategies 1.1 and 1.5; Themes	Review of syllabi and other documents that describe content delivered to teacher candidates.	This is to some extent a qualitative process, but there should be some sort of rubric for adequacy established before the fact.
20	Teacher effectiveness by program disaggregated by student population.	Proposed federal requirement; EPSB Themes	As described by the federal proposed requirement, would amount to some summary (probably a mean) of SGP scores for completers from each program, for each of several demographic groups, such as ethnic minorities or socially disadvantaged students.	This is of questionable validity, but will probably be a federal requirement. Because some of the demographic groups are likely to be quite small, especially when disaggregated by program, these results are likely to be very unstable. Results of Item #1 are likely to be more satisfactory, but will also suffer from problems of instability for small samples. SGP's are really not suitable for this use, as they don't adjust for any of the factors that might affect their magnitude.

21	Technical support for program internal accountability systems	Goal 1, Strategy 1.3	Count of programs with adequate internal accountability systems. Qualitative assessment of instructional validity of existing systems.	Requires initially a survey of existing internal accountability systems. Requires some level of expertise on accountability system development by EPSB staff.
22	Candidate awareness of Code of Ethics	Goal 3, Strategy 3.1	Count of completers who have been exposed to the Code of ethics	Will require a modest additional data collection
23	Completers and alt cert candidates subjected to discipline system	Goal 3, Strategy 3.2	Number of completers/alt cert candidates with complaints; number of completers/alt cert candidates with adjudicated cases (non-dismissed)	Should be limited to completers in the first 5 years of practice
24	Recent completer program satisfaction	Proposed federal requirement; goal 1, Strategy 1.7	Proportion of recent program graduates endorsing "satisfied" or "very satisfied" overall; proportion of school principals endorsing "satisfied" or "very satisfied" overall	System already exists
25	Intern performance	Goal 4, Strategy 4.1	Mean performance score on internship performance record	Requires an improved measurement method

Table 2

Accountability Measure/EPSB goals and strategies cross reference

Goal statement	**Coverage in measure**
Goal 1:	
Strategy 1.1. Maintain regular and rigorous reviews of all program quality indicators.	6, 7, 19
Strategy 1.2. Document and publish information on the quality of each preparation program.	1,2, 9-14
Strategy 1.3. Provide technical assistance to support program improvement.	21
Strategy 1.4. Utilize research to inform program improvements.	18
Strategy 1.5. Review programs to ensure focus on student learning.	19

Strategy 1.6. Maintain a focus on continuous improvement of all preparation programs.	5
Strategy 1.7. Provide accurate and reliable data to support decision making.	18, 24
Goal 2:	
Strategy 2.1. Document every assignment of educators in Kentucky public schools.	None – no relationship to prep program effectiveness
Strategy 2.2. Document the highly qualified status of all Kentucky teachers as required under NCLB.	None – no relationship to prep program effectiveness
Strategy 2.3. Monitor the validity and reliability of teacher and administrator assessments.	15
Strategy 2.4. Document and publish the results of all assessments required of new teachers and new administrators.	11-13
Strategy 2.5. Maintain a focus on continuous improvement of all traditional and alternative route certification procedures and processes.	5,16
Strategy 2.6. Provide accurate and reliable data to support decision making.	18
Goal 3:	
Strategy 3.1. Promote awareness of the EPSB Code of Ethics.	22
Strategy 3.2. Maintain an accurate database of misconduct and character and fitness cases.	23
Strategy 3.3. Present in a timely manner all cases for review by the EPSB.	None – no relationship to prep program effectiveness
Strategy 3.4. Maintain a focus on continuous improvement of all hearing procedures.	None – no relationship to prep program effectiveness
Strategy 3.5. Provide accurate and reliable data to support decision making.	None
Goal 4:	
Strategy 4.1. Develop and utilize reliable measures of teacher effectiveness and student achievement that may be used in evaluation of induction and professional advancement activities.	18, 25
Strategy 4.2. Ensure that every new teacher and principal has a high quality induction experience while demonstrating knowledge and skills that support student learning.	17
Strategy 4.3. Ensure that high quality mentoring and support services are provided for teachers seeking National Board for Professional Teaching Standards certification.	None – no relationship to prep program effectiveness
Strategy 4.4. Ensure that the Continuing Education Option for rank change program maintains appropriate rigor while demonstrating advanced knowledge and skills that support student learning.	None – no relationship to prep program effectiveness
Strategy 4.5. Provide accurate and reliable data to support decision making.	18
Goal 5:	
Strategy 5.1. Maintain a qualified and diverse EPSB workforce.	3,4

Strategy 5.2. Ensure that all personnel are experiencing life-long learning and professional experiences that support their professional growth.	None – no relationship to prep program effectiveness
Strategy 5.3. Seek full funding for all EPSB operations, personnel, and programs through an approved biennial budget request.	None – no relationship to prep program effectiveness
Strategy 5.4. Provide semiannual budget reports to the EPSB.	None – no relationship to prep program effectiveness
Strategy 5.5. Maintain facilities, equipment, and agency technology that support efficient and productive agency operations.	None – no relationship to prep program effectiveness
C. EPSB Themes: Show integration of the following items within the coursework. Choice of formats (matrix, diagram, narrative, etc.) may be used.	
Themes:	3,4,14,20
Additional goals:	
Preparation program district/school support	3
Productivity	4

Table 3

Accountability measures details and sanctions

Item number	**Principle**	**Measure**	**Target**	**Consequences**
1	Teacher quality	Mean effectiveness on K-Prep or SGP adjusted for years post-completion	Within one standard error of state mean	Program review
		Likelihood of student graduation, advanced placement, or other alternative measure given teacher prepared at program, adjusted for years post-completion		Program review
2	Teacher performance appraisal	Mean teacher appraisal adjusted for years post-completion	Within one standard error of state mean	Program review
3	Preparation program attention to district/school	Quality of program plan	Quality is judged to be adequate	Technical assistance

	characteristics	Change in district/school performance attributable to program effort	Detectible change	Technical assistance
4	Productivity	Number of completed candidates in selected content areas compared to criterion	Within one SE of criterion	Possible sanctions
		Total number of completions by content area	Informational only	Report card
5	Candidate screening	Proportion of candidates requiring developmental courses	Close to institutional proportion	Possible sanctions
		Mean candidate GPA	Informational only	Report card
		Mean candidate PPST Math	Informational only	Report card
		Mean candidate PPST Reading	Informational only	Report card
		Mean candidate PPST Writing	Informational only	Report card
		Mean candidate ACT	Informational only	Report card
		Mean graduate candidate GRE	Informational only	Report card
		Number of candidates with dispositions review; number failing admissions because of inadequate dispositions	All candidates screened	Possible sanctions
		Number of applicants subjected to background checks; number failing background checks	All candidates screened	Possible sanctions
6	Accreditation status	Results of most recent accreditation review	Passed review; plan to address exceptions is adequate	Possible sanctions; program closing if egregious and unaddressed
7	Program management	Results of most recent accreditation review	Management procedures are adequate	Possible sanctions; program closing if egregious and unaddressed
8	Candidates completing within a criterion length of time	Proportion of candidates completing within 4 years of admission	> 80%	Technical assistance; possible sanctions
9	Completed candidates who apply for certification within some time horizon	Proportion of completers who apply for certification within 3 years	> 80 %	Technical assistance; possible sanctions

10	Completed candidates who are employed within some time horizon	Proportion of completers employed as teachers within 5 years	> 70%	Technical assistance; possible sanctions
11	Completed candidate Praxis II pass rate (Global Praxis pass rate)	Proportion of completers who passed the relevant Praxis II tests within 3 years of completion	> 80%	Technical assistance; possible sanctions
12	Raw Praxis II pass rate	Proportion of admitted candidates who passed the relevant Praxis II tests within 5 years of completion	> 80%	Technical assistance; possible sanctions
13	Raw Praxis II first-time pass rate	Proportion of completers who passed the relevant Praxis II tests on the first administration within 3 years of completion	> 60%	Technical assistance; possible sanctions
14	Conditional certificates issued	Number/proportion of completers who were issued conditional certificates	<10%	Technical assistance; possible sanctions
		Proportion of conditional completers achieving standard certification in one year	>50%	Technical assistance; possible sanctions
15	Review of assessments	Evidence that every assessment has been reviewed annually	All assessments reviewed	N/A
16	Preparation productivity by type of program (e.g., alternative or traditional)	Number of completers by type of program in each content area	Informational only	Report card
		Proportion of alternative candidates completing program	> 80%	Technical assistance; possible sanctions
		Number of alternative candidates transferring to a traditional program in same institution	< 10%	Technical assistance; possible sanctions
		Number of alternative candidates transferring to a program in another institution	< 10%	Technical assistance; possible sanctions
17	Program management of clinical/induction support	Number of student teachers with inadequate mentorship	0	Technical assistance; possible sanctions
		Number of alternative candidates with inadequate mentorship	0	Technical assistance; possible sanctions
18	Annual review of the accountability system	Evidence accountability system was reviewed (annual report)	Report satisfactorily completed	N/A
		Unintended consequences identified	N/A	N/A
		Existing measures requiring adjustment	N/A	N/A

		Possible new measures identified	N/A	N/A
19	Review of preparation program content	Required content in place	All required content in place	Technical assistance; possible sanctions
20	Teacher effectiveness by program disaggregated by student population.	Program completer mean SGP disaggregation by ethnicity	Evidence of progress	Technical assistance
		Program completer mean SGP by gender	Evidence of progress	Technical assistance
		Program completer mean SGP by economically disadvantaged	Evidence of progress	Technical assistance
21	Technical support for program internal accountability systems	Internal accountability systems present	Present in all institutions	Technical assistance
		Internal accountability systems are instructionally valid	Demonstrated validity	Technical assistance
22	Candidate awareness of Code of Ethics	Number of completers exposed to code of ethics	All completers exposed	Technical assistance
23	Completers and alt cert candidates subjected to discipline system	Number/proportion of completers (within 5 years of completion) receiving complaints	Proportion within one SE of state average	Program review
		Number/proportion of appropriate cases adjudicated (not dismissed)	Informational only	Report card
24	Recent completer program satisfaction	Proportion of recent program graduates endorsing “satisfied” or “very satisfied” overall	>75%	Report card
		Proportion of school principals endorsing “satisfied” or “very satisfied” overall	>75%	Report card
25	Intern performance	Mean performance score on internship performance record	Within one SE of state mean	Technical assistance; program review

References

Abadie A & Imbens G (2011). Matching on the estimated propensity score.

Abedi J, Goldschmidt P, Gong B, Gottlieb M, Ortiz A, Pedraza P, Pellegrino J, Roschewski P & Stack J (2007). Assessment and accountability for improving schools and learning: Principles and recommendations for federal law and state and local systems. Forum on Educational Accountability.

Ahearn E (2000). Educational accountability: A synthesis of the literature and review of a balanced model of accountability. Alexandria, Va.: Final Report Deliverable #2-2.2a Under Cooperative Agreement No. H159K70002 to the U.S. Department of Education, Office of Special Education Programs by the National Association of State Directors of Special Education.

Alliance for Excellent Education (2011). Waiving away high school graduation rate accountability? Washington, D.C.: Author, policy Brief, January 2012.

American Evaluation Association (2006). Public statement: Educational accountability.

Amrein-Beardsley A (2008). Methodological concerns about the Education Value-Added Assessment System. Educational Researcher, 37(2):65-75.

Amrein-Beardsley A & Collins C (2012). The SAS Education Value-Added Assessment System (SAS EVAAS) in the Houston Independent School District (HISD): Intended and unintended consequences. Education Policy Analysis Archives 20(12).

Amrein A & Berliner D (2002). High-stakes testing, uncertainty, and student learning. Education Policy Analysis Archives, 10(18). Retrieved April 2012 from http://epaa.asu.edu/epaa/v10n18/.

Ananda S & Rabinowitz S (2003). Building a workable accountability system. Wested knowledge brief.

Armour-Garb A (2008). Structural problems in educational accountability. The Nelson A. Rockefeller Institute of Government, Education policy brief.

Balter D & Duncomb W (2008). Recruiting highly qualified teachers: Do district recruitment practices matter? Public Finance Review, 36:33-62.

Baker E (2003). Multiple measures: Toward tiered systems. Educational Measurement Issues and Practice. Summer 2003:13-17.

Baker E (2005). Improving accountability models by using technology-enabled knowledge systems (TEKS). University of California, Los Angeles , National Center for Research on Evaluation, Standards, and Student Testing (CRESST) CSE Report 656

Baker E, Barton P, Darling-Hammond L, Haertel E, Ladd H , Linn R, Ravitch D, Rothstein R, Shavelson R & Shepard L (2010). Problems with the use of student test scores to evaluate teachers. Economic Policy Institute, EPI Briefing Paper # 278.

Ballou D (2004). Rejoinder. Journal of Educational and Behavioral Statistics 29(1) Value-Added Assessment Special Issue 131-134.

Bathgate K, Colvin R & Silva E (2011). Striving for student success: A model of shared accountability. Washington, D.C.: Education Sector.

Betebenner D (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. Dover, New Hampshire: The National Center for the Improvement of Educational Assessment.

Bettebenner D & Linn R (2009). Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability. Paper presented at the Center for K-12 Assessment and Performance Management exploratory seminar: Measurement Challenges Within the Race to the Top Agenda, ETS.

Biesta G (2008). Good education in an age of measurement: on the need to reconnect with the question of purpose in education. Educational Assessment, Evaluation and Accountability 21:33–46.

Blanton L, Sindelar P, Correa V, Hardman M, McDonnell J & Kuhel K (2003). Conceptions of beginning teacher quality: Models for conducting research. Center on Personnel Studies in Special Education.

Board of Testing and Assessment of the National Research Council (2009). Letter report to the U.S. Department of Education on the Race to the Top Fund. Retrieved April 2012 from http://www.nap.edu/catalog.php?record_id=12780

Bovens M (2010). Two concepts of accountability: accountability as a virtue and as a mechanism, West European Politics 33: 946–967.

Boyd D, Lankford H, Loeb S & Wyckoff J (2008). The impact of assessment and accountability on teacher recruitment and retention: Are there unintended consequences? Public Finance Review, 36(1):88-111.

Boyd D, Grossman P, Lankford H, Loeb S & Wyckoff J (2005). The draw of home: Teachers' preferences for proximity disadvantage urban schools. Journal of Policy Analysis and Management 24(1) 113-132.

Briggs D & Weeks J (2009). The impact of vertical scaling decisions on growth interpretations. Educational Measurement: Issues and Practice 28(4):pp. 3–14.

Bruns B, Filmer D & Patrinos (2011). Making schools work: New evidence on accountability reforms. Washington, D.C.: The World Bank.

Bryan S & Hofmann B (2007). Transparency and accountability in Africa's extractive industries: The role of the legislature. Washington, D.C.: National Democratic Institute for International Affairs.

Bryk A (2003). No Child Left Behind, Chicago style. In Peterson P & West M (2003). No child left behind? The politics and practice of school accountability. Washington, D.C.: Brookings Institution press.

Bublick E (2003). Comparative fault to the limits. Vanderbilt Law Review 56(4): 977-1045

Buckley K & Marion S (2011). A survey of approaches used to evaluate educators in non-tested grades and subjects. Downloaded March 2013 from http://www.google.com/url?sa=t&rct=j&q=buckley%20marion%20a%20survey%20of%20approaches&source=web&cd=4&cad=rja&ved=0CD4QFjAD&url=http%3A%2F%2Fwww.nciea.org%2Fpublications%2FSummary%2520of%2520Approaches%2520for%2520non-tested%2520gradesKBSM2011.pdf&ei=QoxMUeinHK-j4APZ4oDYCA&usg=AFQjCNE85gC1YpbHmNK6yMUd4i9a4aNR7A

Campbell R, Kyriakides L, Muijs R & Robinson W (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. Oxford Review of Education 29(3) 347-362.

Carnegie Forum on Education and the Economy (1986). A nation prepared: teachers for the 21st century. Report of the Task Force on Teaching as a Profession.

Carnoy M & Loeb S (2002). Does external accountability affect student outcomes? A Cross-State Analysis. Educational Evaluation and Policy Analysis, 24(4) 305-331

Center for Reform of School Systems (2003). Urban school district accountability systems. Author.

Chassin M, Loeb J, Schmaltz S & Wachter R (2010). Accountability measures — using measurement to promote quality improvement. New England Journal of Medicine 363:683-688

Chater N, Oaksford M, Hahn U & Heit E (2009). Inductive logic and empirical psychology. Handbook of the History of Logic, vol. 19.

Chen W, Mason S, Staniszewski C, Upton A & Valley M (2012). Assessing the quality of teachers' teaching practices. Educational Assessment, Evaluation, and Accountability 24:25–41.

Chester M (2005). Making valid and consistent inferences about school effectiveness from multiple measures. Educational Measurement: Issues and Practice, Winter 2005:40-52.

Chetty R, Friedman J, & Rockoff J (2012). The Long-term impacts of teacher value-added and student outcomes in adulthood. National Bureau of Economic Research NBER Working Paper No. 17699.

Clark D & See E (2011). The impact of tougher education standards: Evidence from Florida. Economics of Education Review 30:1123– 1135.

Clotfelter C, Ladd H & Vigdor J (2006). Teacher-student matching and the assessment of teacher effectiveness. National Bureau of Economic Research Working Paper 11936.

Clotfelter C, Ladd H, & Vigdor J (2007a). How and why do teacher credentials matter for student achievement? National Center for Analysis of Longitudinal Data In Education Research, Working Paper 2.

Clotfelter C, Ladd H & Vigdor J (2007b). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. National Center for Analysis of Longitudinal Data In Education Research, Working Paper 11.

Coca-Perraillon M (2006). Matching with propensity scores to reduce bias in observational studies. Paper presentation at Northeast SAS Users' Group 2006, September 17 - 10, 2006 Philadelphia, PA.

Coggshall J (2007). Communication framework for measuring teacher quality and effectiveness: Bringing coherence to the conversation. National Comprehensive Center for Teacher Quality.

Coggshall J, Bivona L & Reschly (2012). Evaluating the effectiveness of teacher preparation programs for support and accountability. National Comprehensive Center for Teacher Quality research brief. Accessed February 2013 from: http://www.tqsource.org/publications/TQ_RandP_BriefEvaluatingEffectiveness.pdf.

Columbia University Teacher's College (2009). Has NCLB improved teacher quality? Accessed February 2013 from http://www.tc.columbia.edu/news.htm?articleID=6989.

Council for the Accreditation of Teacher Preparation (CAEP) (2013). Draft recommendations for the CAEP Board. Washington, D.C.: author. Accessed February 2013 from: http://caepnet.files.wordpress.com/2013/02/draft_standards.pdf#page=1&zoom=auto,0,792

Crowe, E (2011). Analyzing State Strategies for Ensuring Real Accountability and Fostering Program Innovation. Center for American Progress.

D'Agostino J, Welsh M, & Corson N (2007). Instructional Sensitivity of a State's Standards-Based Assessment. Educational Assessment *12*(1) 1–22.

Damon W (2007). Dispositions and teacher assessment : The need for a more rigorous definition. *Journal of Teacher Education* 58: 365-369.

Dearborn M (2009). Enterprise liability: Reviewing and revitalizing liability for corporate groups. California Law Review 97: 195-261.

Dee T & Jacob B (2009). The impact of no child left behind on student achievement. National Bureau of Economic Research, NBER Working Paper 15531.

Dorn S (1998). The political legacy of school accountability systems. Education Policy Analysis Archives, 1(6)

Downes T & Figlio D (1998). School finance reforms, tax limits, and student performance: Do reforms level up or dumb down? Institute for Research on Poverty Discussion Paper no. 1142-97.

Downey D, Hippel P & Broh B (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. American Sociological Review, 69(5):613-635.

Draut K (2012, March). Overview of next-generation learners: Part 2 (growth). Powerpoint presentation. Frankfort, KY: Kentucky Department of Education, Office of Assessment and Accountability.

Dunn J & Allen J (2009). Holding schools accountable for the growth of nonproficient Students: Coordinating measurement and accountability. Educational Measurement: Issues and Practice 28(4):27–41.

Dunn D (2003). Accountability, democratic theory, and higher education. Educational Policy, 17(1): 60-79.

Earley P (2001). Title II requirements for schools, colleges, and departments of education. ERIC Digest. ED460124.

Edwards M (2011). 'Shared accountability' in service delivery: concepts, principles and the Australian experience. Paper presented at the UN Committee of Experts on Public Administration Vienna Meeting July 2011.

Ellermann D (2005). The two institutional logics: Exit-oriented versus commitment-oriented institutional designs. International Economic Journal, 19(2):147–168.

Ellett C & Teddlie C (2003). Teacher evaluation, teacher effectiveness and school effectiveness: Perspectives from the USA. Educational Assessment, Evaluation and Accountability 17(1) 101-128.

Elmore R (2003). A plea for strong practice. The Challenges of Accountability 6(3): 6-10.

Ewell P (2009). Assessment, accountability, and improvement: Revisiting the tension. National Institute for Learning Outcomes Assessment.

Ewell P, Boeke M, & Zis S (2008). State policies on student transitions: Results of a fifty-state inventory. National Center for Higher Education Management Systems (NCHEMS).

Ewing, J (2011). Mathematical intimidation: Driven by the Data. Notices of the AMS 58(5).

Ferrara S, Svetina D, Skucha S, & Davidson A(2011). Test development with performance standards and achievement growth in mind. Educational Measurement: Issues and Practice 30(4):3–15.

Figlio D & Kenny L (2007). Individual teacher incentives and student performance. Journal of Public Economics 91: 901–914.

Figlio D & Loeb S (2011). School accountability. In Hanushek E, Machin S & Woessmann L (eds.) Handbooks in Economics, Volume 3, The Netherlands: North-Holland, 2011, pp. 383-421.

Figlio D & Lucas M (2004). What's in a grade? School report cards and the housing market. The American Economic Review, 94(3):591-604.

Floden R (2012). Teacher Value Added as a Measure of Program Quality: Interpret With Caution. Journal of Teacher Education 63(5) 356–360.

Frankel A, Leonard M & Denham C (2006). Fair and just culture, team behavior, and leadership engagement: The tools to achieve high reliability. Health Services Research 41(4 Pt 2): 1690–1709.

Fuhrman S (2003). Redesigning accountability systems for education. Consortium for Policy Research in Education CPRE Policy Brief RB-38.

Furger F (1997). Accountability and systems of self-governance: The case of the maritime industry. Law and Policy 19(4): 445-476.

Gain Working Group E (2004). A roadmap to just culture: Enhancing the safety environment. Flight Ops/ATC Ops Safety Information Sharing Working Group of the Global Aviation Information Network

Gilboa I & Samuelson L (2012). Subjectivity in inductive inference. Theoretical Economics 7:183–215.

Ginsberg R & Whaley D (2006). The disposition on dispositions. The Teacher Educator 4(4). 269-275.

Goe L, Bell C & Little O (2008). Approaches to evaluating teacher effectiveness: A research synthesis. National Comprehensive Center for Teacher Quality.

Goe L & Croft A (2009). Methods of evaluating teacher effectiveness. National Comprehensive Center for Teacher Quality Research-to-Practice Brief.

Goldhaber D, Brewer D & Anderson D (1999). A three-way error components analysis of educational productivity. Education Economics, 7(3):199-208.

Goldhaber D & Evans D (2007). The importance of methodology in teasing out the effects of school resources on student achievement. Center on Reinventing Public Education, Working Paper 2007-5.0.

Goldhaber D & Hansen M (2012). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. National Center for Analysis of Longitudinal Data In Education Research, Working Paper 73.

Goldhaber D, Koedel C, Loeb S & Sass T (2012). CALDER Conversations Topic 1: Evaluating teacher training programs. Retrieved from www.caldercenter.org/calder-conversations-tpps.cfm

Goldhaber D & Liddle (2012). The gateway to the profession: Assessing teacher preparation programs based on student achievement. National Center for Analysis of Longitudinal Data in Education Research, Working Paper 65.

Gong B (2002). Designing school accountability systems: Towards a framework and process. State Collaborative on Assessment and Student Standards (SCASS) Accountability Systems and Reporting (ASR) Consortium.

Gunzenhauser M & Hyde A (2007). What is the value of public school accountability? Educational Theory, 57(4):489-507.

Hanushek E (2003). The confusing world of educational accountability. The National Tax Journal, 54(2):365-384.

Hanushek E & Raymond M (2003). Lessons about the design of state accountability systems. In Peterson P & West M (2003). No child left behind? The politics and practice of school accountability. Washington, D.C.: Brookings Institution press.

Hanushek E & Raymond M (2005). Does school accountability lead to improved student performance? Journal of Policy Analysis and Management, 24(2):297-327.

Heinrich C, Maffioli A, & Vázquez G (2010). A primer for applying propensity-score matching. Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank Technical Notes

Heneman H, Milanowski A, Kimball S & Odden A (2006). Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay. University of Pennsylvania Graduate School of Education, CPRE Policy Brief RB-45.

Henry G, Thompson C, Bastian K, Fortner C, Kershaw D, Purtell K, & Zulli R (2011). Does teacher preparation affect student achievement? Manuscript submitted to Education Finance and Policy, February 7, 2011.

Hess F (2003). Refining or retreating? High-stakes accountability in the states. In Peterson P & West M (2003). No child left behind? The politics and practice of school accountability. Washington, D.C.: Brookings Institution press.

Hibpshman T (2004). Distribution of teachers by institution. Frankfort, Kentucky: Education Professional Standards Board. Unpublished manuscript.

Hibpshman T (2006). Development of a replacement for the QPI. Frankfort, Kentucky: Education Professional Standards Board, unpublished manuscript.

Hibpshman T (2007). Distribution of Teachers in Kentucky. Frankfort, Kentucky: Education Professional Standards Board, unpublished manuscript.

Hibpshman (2010). Email to Phillip Rogers, Executive Director of the Education Professional Standards Board.

Hibpshman T (2012). A review of teacher effectiveness research. Unpublished manuscript.

Ho A, Lewis D & Farris J (2009). The dependence of growth-model results on proficiency cut scores. Educational Measurement: Issues and Practice, 28(4):15–26.

Hoffmann D, Jacons R & Baratta J (1993). Dynamic criteria and the measurement of change. Journal of Applied Psychology, 78(2):194-204.

Hollingshead L & Childs R (2011). Reporting the percentage of students above a cut score: The effect of group size. Educational Measurement: Issues and Practice, 30(1):36–43.

Jackson C & Bruegmann E (2009). Teaching students and teaching each other: The importance of peer learning for teachers. American Economic Journal: Applied Economics 1(4) 85-108.

Jackson E & Page M (2013). Estimating the distributional effects of education reforms: A look at Project STAR. Economics of Education Review 32: 92–103.

Jacob B (2003). A closer look at achievement gains under high-stakes testing in Chicago. In Peterson P & West M (2003). No child left behind? The politics and practice of school accountability. Washington, D.C.: Brookings Institution press.

Jacob B & Lefgren L (2007). Can principals identify effective teachers? Evidence on subjective performance evaluation in education.

Jensen P & Rasmussen A (2011). The effect of immigrant concentration in schools on native and immigrant children's reading and math skills. Economics of Education Review, 30: 1503– 1515.

Jin G & Leslie P (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. The Quarterly Journal of Economics, 118(2):409-451.

The Joint Commission (2010). Facts about accountability measures. Downloaded January 2013 from http://www.google.com/url?sa=t&rct=j&q=facts%20about%20accountability%20measures&source=web&cd=4&cad=rja&ved=0CEEQFjAD&url=http%3A%2F%2Fwww.jointcommission.org%2Fassets%2F1%2F18%2FFacts_about_Accountability_measures.pdf&ei=whsgUYe4I7Ko0AHM_YGYDQ&usg=AFQjCNH-FIIy0B50nFmN7UOnTcS8udO7hw

Kane T & Staiger D (2002). Improving school accountability measures. Cambridge, Massachusetts: National Bureau of Economic Research NBER Working Paper 8156.

Kane T & Staiger D (2008). Estimating teacher impacts on student achievement: An experimental evaluation. National Bureau of Economic Research, NBER Working Paper 14607.

Kane T, Staiger D & Geppert J (2002). Randomly accountable. Education Next, Spring 2002:57-61.

Kelcey B (2011). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. Educational Evaluation and Policy Analysis 33: 458-482.

Kentucky Education Professional Standards Board (2012a). EPSB 2011-12 Goals and Strategies. Frankfort, Kentucky: author. Accessed March 2013 at: http://www.epsb.ky.gov/boardinfo/mission.asp

Kentucky Education Professional Standards Board (2012b). Kentucky Program Guidelines - Initial and Advanced Program. Frankfort, Kentucky: author. Accessed March 2013 at: http://www.kyepsb.net/teacherprep/programguidelines.asp

Kentucky Department of Education (2011). EOC Constructed-Response Outline. Frankfort, KY: author.

Kentucky Department of Education (2012). K-PREP Blueprint. Frankfort, KY: author.

Kentucky Department of Education (2013). Professional Growth and Effectiveness System Winter Summit. Frankfort, Kentucky: author. Accessed February 2013 from: https://skydrive.live.com/?cid=27a47713f201df2c&id=27A47713F201DF2C!1166&authkey=!ALhmA9EqG0EjAG8#!/?cid=27a47713f201df2c&id=27A47713F201DF2C!1166&authkey=!ALhmA9EqG0EjAG8

Kentucky Legislative Research Commission (2012). 703 KAR 5:225. School and district accountability, recognition, support, and consequences.

Kim S (2004). Accountability and school reform in the U.S. public school system. Downloaded February 2013 from https://pantherfile.uwm.edu/kim/www/papers/Accountability_Final.pdf

Kimball S (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. Educational Assessment, Evaluation and Accountability 16(4) 241-268.

Kinsler J (2009). Beyond levels and growth: Estimating teacher value-added and its persistence. Paper presented at the 34th Annual AEFP Conference.

Koedel C & Betts J (2009). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. Downloaded June 2012 from http://economics.missouri.edu/working-papers/2009/wp0902_koedel.pdf

Koretz D (2006). The pending reauthorization of NCLB: An opportunity to rethink the basic strategy. Invited Paper for Civil Rights Project/ Earl Warren Institute Roundtable discussion on the reauthorization of NCLB Washington, D.C. November 16, 2006.

Koretz D (2010). Implications of current policy for educational measurement. Educational Testing Service. Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, Center for K – 12 Assessment & Performance Management.

Koretz D & Barron S (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS). Rand Corporation.

Krynski T & Tenenbaum J (2007). The role of causality in judgment under uncertainty. Journal of Experimental Psychology – General, 136(3) 430-450.

Kuh G (2007). Risky business: Promises and pitfalls of institutional transparency. Change, September/October 2007:30-35.

Kukla-Acevedo S, Streams M & Toma E (2011). Can a single performance metric do it all? A Case Study in Education Accountability. The American Review of Public Administration 42(3) 303–319.

Kyburg H & Teng C (2001). Uncertain inference. Cambridge, UK: Cambridge Press.

Ladd H (2001). School-based educational accountability systems: The promise and the pitfalls. National Tax Journal 54(2):385-400.

Ladd H & Walsh R (2002). Implementing value-added measures of school effectiveness: getting the incentives right. Economics of Education Review, 21:1–17.

Linn L (2003) Accountability: Responsibility and reasonable expectations. Educational Researcher, 32(7): 3-13.

Liu O (2011). Outcomes assessment in higher education: Challenges and future research in the context of voluntary system of accountability. Educational Measurement: Issues and Practice Fall2011, 30(3):2–9.

Loeb S & Strunk K (2007). Accountability and local control: Response to incentives with and without authority over resource generation and allocation. Education Finance and Policy 2007 10-39

Lubienski S & Crane C (2010) Beyond free lunch: Which family background measures matter? Education Policy Analysis Archives*, 18*(11). Retrieved May 2012, from http://epaa.asu.edu/ojs/article/view/756.

Macdonald T & Macdonald K (2006). Non-electoral accountability in global politics: Strengthening democratic control within the global garment industry. The European Journal of International Law Vol. 17(1): 89-119.

Mangiante E (2011). Teachers matter: Measures of teacher effectiveness in low-income minority schools. Educational Assessment, Evaluation and Accountability 23:41–63.

Manna P (2010). The No Child Left Behind Act and educational accountability in the United States. Paper prepared for the Conference on Understanding and Evaluating New Accountability Regimes: Canada in Comparative Perspective, University of Toronto, Canada, February 5-6, 2010.

Mathios A (2000). The impact of mandatory disclosure laws on product choices: An analysis of the salad dressing market. Journal of Law and Economics, 43(2): 651-678.

Matsumura L, Garnier H, Pascal J & Valdés R (2002). Measuring Instructional Quality in Accountability Systems: Classroom Assignments and Student Achievement. Educational Assessment *8*(3), 207–229.

Mason P (2010) Assessing difference; examining Florida's initial teacher preparation programs and exploring alternative specifications of value-added models. MPRA Paper No. 27903.

McClellan M & Staiger D (1999). The quality of health care providers. National Bureau of Economic Research, NBER Working Paper # 7327.

Micklewright J, Schnepf S, Silva P (2012). Peer effects and measurement error: The impact of sampling variation in school survey data (evidence from PISA). Economics of Education Review, 31: 1136– 1142.

Milanowski A, Kimball S & White B (2004). The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites. Consortium for Policy Research in Education, CPRE-UW Working Paper Series TC-04-01.

Muijs D (2006). Measuring teacher effectiveness: Some methodological reflections. Educational Research and Evaluation 12(1) 53 – 74.

Mulgan R (2000). 'Accountability': An ever-expanding concept? Public Administration 78(3):555–573.

Murray F (2007). Disposition: A superfluous construct in teacher education. Journal of Teacher Education 58: 381-387.

Nace D & Gartland J (2011). Providing accountability: Accountable care concepts for providers. McKessen Relay Health

National Academy of Education (2009). Standards, assessment, and accountability. Author: education policy white paper.

The National Commission on Excellence in Education (1983). A Nation at Risk: The Imperative for Educational Reform. A Report to the Nation and the Secretary of Education United States Department of Education by The National Commission on Excellence in Education. Accessed February 2013 from: http://datacenter.spps.org/uploads/SOTW_A_Nation_at_Risk_1983.pdf

Neal D & Schanzenbach D (2010). Left behind by design: Proficiency counts and test-based accountability. The Review of Economics and Statistics, 92(2):263–283.

Newmann F, King M & Rigdon M (1997). Accountability and school performance: Implications from restructuring schools. Harvard Educational Review, 67(1):41-69.

Nichols P, Meyers J & Burling K (2009). A framework for evaluating and planning assessments intended to improve student achievement. Educational Measurement: Issues and Practice 28(3):14–23.

Nichols P, Twing J, MuellerC & O'Malley K (2010). Standard-setting methods as measurement processes. Educational Measurement: Issues and Practice, 29(1):14–24.

Noell G (2007). Value added assessment of teacher preparation in Louisiana: 2004 – 2006. Louisiana State University

Noell G & Burns J (2006). Value-added assessment of teacher preparation : An illustration of emerging technology. Journal of Teacher Education 57, 37.

Normand S, Glickman M & Gatsonis C (1997). Statistical methods for profiling providers of medical Care: issues and applications. Journal of the American Statistical Association, 92(439):803-814.

O'Day J (2002). Complexity, accountability, and school improvement. Harvard Educational Review, 72(3)

Omar, M (2003). Statistical process control charts for measuring rating consistency over time. Dhahran, Saudi Arabia: King Fahd University of Petroleum and Minerals, Department of Mathematics and Statistics.

Oruko L, Randall I, Bwalya M, Kisira S and Wanzala M (2011). Mutual accountability framework for the comprehensive Africa agriculture development programme. African union, NEPAD Planning and Coordinating Agency.

Osborne C, von Hippel P, Lincove J, Mills N & Bellows L (2013). The Small and Unreliable Effects of Teacher Preparation Programs on Student Test Scores in Texas. Paper presented at the AFP Conference, 2013.

Parsons L (2004). Performing a 1:N case-control match on propensity score. Paper presented at SUGI 29, Montreal, May 9-12 2004.

Pellegrino J (2010). The Design of an Assessment System for the Race to the Top: A Learning Sciences Perspective on Issues of Growth and Measurement. Educational Testing Service.

Peterson P & West M (2003). No child left behind? The politics and practice of school accountability. Washington, D.C.: Brookings Institution press.

Plecki M, Elfers A & Nakamura Y (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value added measures. Journal of Teacher Education 63(5) 318 –334.

Polikoff M (2010). Instructional sensitivity as a psychometric property of assessments. Educational Measurement: Issues and Practice, 29(4): 3–14.

Reback R (2007). Teaching to the rating: School accountability and the distribution of student achievement. Journal of Public Economics, 92:1394–1415.

Reed B (2005). Accountability in a shared-services world. Future challenges for e-government: Collective accountability. Discussion paper no. 10.

Rivkin S, Hanushek E & Kain J (2005). Teachers, schools, and academic achievement. Econometrica, 73(2): 417-458.

Rockoff J & Turner L (2010). Short run Impacts of accountability on school quality.

Romzek B & Dubnick M (1987). Accountability in the public sector: Lessons from the Challenger tragedy. Public Administration Review, 47(3): 227-238.

Rothstein R (2008b). Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education. National Center on Performance Incentives.

Rubin D & Thomas N (1995). Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association 95(4) 573-585.

Sanders W & Horn P (1994). The Tennessee value-added assessment system. Mixed methodology and educational assessment. Journal of Personnel Evaluation in Education, 8(1):299-311.

Sansani S (2011). The effects of school quality on long-term health. Economics of Education Review, 30:1320– 1333.

Sass T (2008). The stability of value-added measures of teacher quality and implications for teacher compensation policy. National Center for Analysis of Longitudinal Data In Education Research, Research Brief 4.

Schafer W, Coverdale B, Luxenberg H & Jin Y (2011). Quality control charts in large-scale assessment programs. Practical Assessment, Research, and Evaluation, 16(15).

Scholle S, Sampsel S & Davis N (2009). Quality of child health care: Expanding the scope and flexibility of measurement approaches. National Committee for Quality Assurance Issue Brief.

Seltzer M, Choi K & Thum Y (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. Educational Evaluation and Policy Analysis, 25(3): 263-286.

Sims D (2013). Can failure succeed? Using racial subgroup rules to analyze the effect of school accountability failure on student performance. Economics of Education Review 32:262–274.

Skrla L, Scheurich J, Johnson J, Hogan D, Koschoreck J & Smith P (2000). A report on systemic school success in four Texas school districts serving diverse student populations. University of Texas at Austin, Charles A Dana Center.

Sturman M, Cheramie R & Cashen L (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test–retest reliability of employee job performance ratings. Journal of Applied Psychology 90(2):269–283.

Supovitz J (2010). Is high-stakes testing working? Philadelphia, PA: University of Pennsylvania Graduate School of Education.

Swanson C & Stevenson D (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. Educational Evaluation and Policy Analysis 24(1) 1-27.

Torre C & Voyce C (2007). Shared accountability: An organic approach. In: Després, Blane (Editor). (2007). Systems Thinkers in Action: A Field Guide for Effective Change Leadership in Education. Rowman & Littlefield Publishers.

Tyack D & Cuban L (1995). Tinkering toward Utopia. Cambridge, Massachusetts: Harvard University.

U.S. Department of Education (USDOE) (2013). Title II homepage. Accessed April 2013 at https://title2.ed.gov/

Walters-Parker K (2012). Personal communication.

Ward C (2000). GI Forum v. Texas Education Agency: Implications for state assessment programs. Applied Measurement in Education, 13(4):419–426.

Wei X & Haertel E (2011). The effect of ignoring classroom-level variance in estimating the generalizability of school mean scores. Educational Measurement: Issues and Practice 30(1):13–22.

Weimer D (2001). School performance and housing values: Using non-contiguous district and incorporation boundaries to identify school effects.

Wilensky J, Galvin K, & Pascoe D (2004). Educational accountability systems: Motivation or discrimination? A survey of the legal theories used to challenge and defend educational accountability systems. Paper prepared for the conference: "50 Years after Brown: What Has Been Accomplished and What Remains to Be Done?" Kennedy School of Government, Harvard University, April 23-24, 2004

Williams S, Schmaltz S, Morton D, Koss R, & Loeb J (2005). Quality of care in U.S. hospitals as reflected by standardized measures, 2002–2004. New England Journal of Medicine 353(3): 255 – 264.

Woodall W, Adams B & Benneyan J (2011). The use of control charts in healthcare. In Faltin F, Kenett R & Ruggeri F, eds., Statistical Methods in Healthcare, Wiley, 2011.

Wong M, Cook T & Steiner P (2011). No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series. Evanston, Illinois: Northwestern University Institute for Policy Research Working Paper Series WP-09-11.

