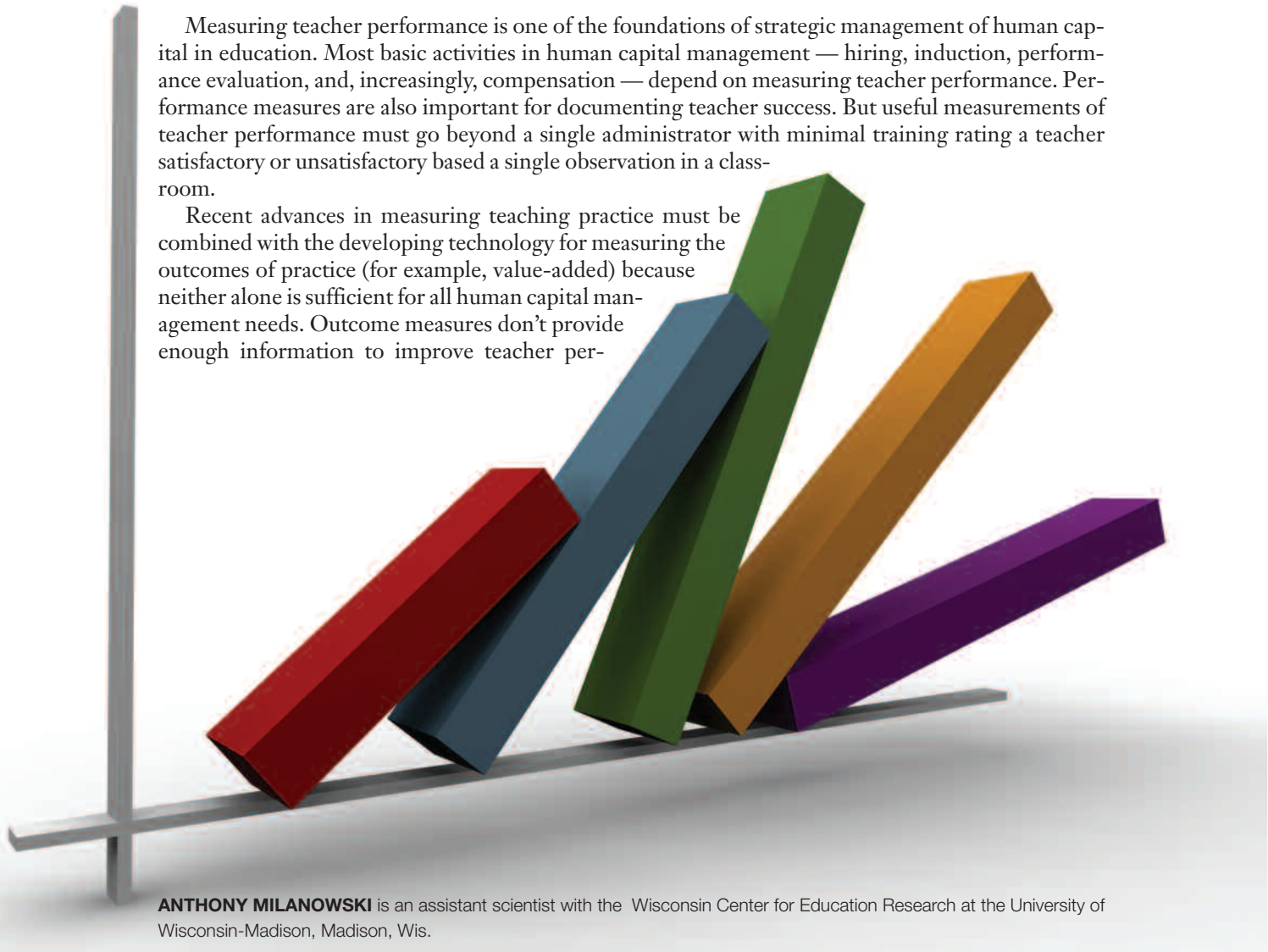
Strategic Measures of Teacher Performance

Moving from our current haphazard system of teacher evaluation to a more systematic approach to evaluation is a cohesive strategy for improving the performance of our schools.

By Anthony Milanowski

Measuring teacher performance is one of the foundations of strategic management of human capital in education. Most basic activities in human capital management — hiring, induction, performance evaluation, and, increasingly, compensation — depend on measuring teacher performance. Performance measures are also important for documenting teacher success. But useful measurements of teacher performance must go beyond a single administrator with minimal training rating a teacher satisfactory or unsatisfactory based a single observation in a classroom.

Recent advances in measuring teaching practice must be combined with the developing technology for measuring the outcomes of practice (for example, value-added) because neither alone is sufficient for all human capital management needs. Outcome measures don't provide enough information to improve teacher per-



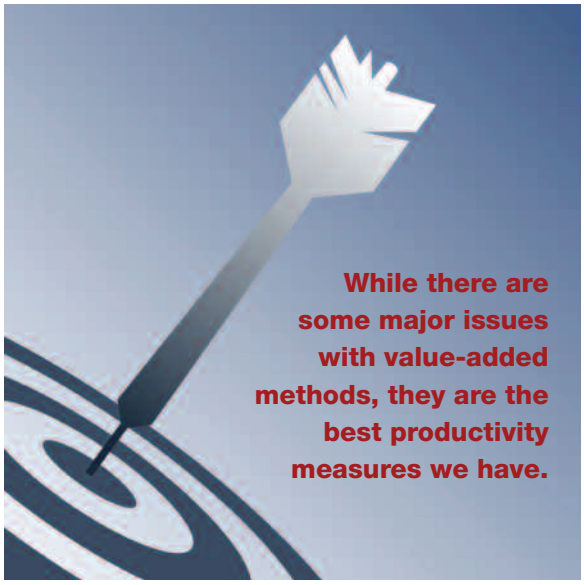
Thinkstock/Hemera

ANTHONY MILANOWSKI is an assistant scientist with the Wisconsin Center for Education Research at the University of Wisconsin-Madison, Madison, Wis.

formance. Likewise, instructional practice measures that aren't linked to effects on learning are likely to lose their rigor and relevance. At this stage of the policy debate, many stakeholders will accept a practice measure only if it can be linked to the "bottom line": student outcomes.

MEASURING PRACTICE

Measuring teaching practice begins by translating a vision of effective instruction into a model that makes explicit what competent performance is. This model would summarize the teaching behaviors and related skills that constitute effective teaching, including instructional planning, classroom management, delivering instruction, and what teachers need to know and be able to do to implement particular state or district instructional strategies. The model then provides a template for aligning all state and district



measurements and practices so they work together to acquire, develop, motivate, and retain teachers with the requisite competencies.

Developing a teacher competency model doesn't require reinventing the wheel. Designers can start with state teaching standards, standards promulgated by such national organizations as INTASC and the National Board for Professional Teaching Standards and by such frameworks as Charlotte Danielson's Framework for Teaching or the New Teacher Center's Continuum of Teacher Development. Beginning with an existing model captures those aspects of teaching that are similar across states and districts. Designers would then add competencies that reflect a particular state or district vision and those needed to support local instructional initiatives and strategies. Several districts (e.g., Chicago and Cincinnati) and states (e.g., Idaho and Delaware) have modified Danielson's Framework for their teacher evaluation systems. Allan Odden and Marc Wallace (2008) present another modification to this framework, with 15 standards addressing planning, classroom management, delivering instruction, reflection on teaching, collaboration with colleagues, and communication with families.

Another approach might be to use the Classroom Assessment Scoring System (CLASS) for competencies related to classroom management, student engagement, and teacher-student interactions, and then add specific competencies related to instructional planning, student assessment, and district instructional strategies. The Measuring Effective Teaching project, supported by the Bill & Melinda Gates Foundation, is testing the use of CLASS combined with a more subject-specific assessment (e.g., in mathematics and language arts).

MEASUREMENT SYSTEMS

Three measurement systems are needed to assess teaching practice: 1) observations of classroom practice for use in periodic formal teacher evaluation, 2) teaching "work samples" or performance assessments for such decisions as granting tenure or movement on a career ladder, and 3) classroom walkthroughs that provide information for everyday performance management. All three are needed to ensure complete coverage of important competencies and to allow for different uses.

Classroom observations. To be useful, classroom observations must be based on a measurement system that promotes reliability and validity. An observation system should include:

- Multiple competency levels defined by rating scales or rubrics that provide concrete examples of the levels. Rubrics guide evaluators in making more reliable decisions, provide teachers with concrete descriptions of what good performance looks like, and communicate performance expectations.
- Procedures for gathering and evaluating evidence that are clearly delineated so they can be implemented uniformly.
- A focus on aspects of instructional practice that can be observed in a typical instructional period, such as student behavior management, use of instructional time, rapport with students, student engagement, and lesson adjustment. Though observations shouldn't ignore teacher content knowledge and application of content-specific pedagogy, these are hard to observe in a limited set of classroom observations, especially by an administrator who has no background in a

content area. In-depth assessment of these areas should be done using teaching work samples (discussed further below).

- Trained observers. Training should use videos or live practice observations to help observers understand how to recognize levels of competency and apply the rubrics. Observers should be assessed on their ability to apply the rubrics. The Cincinnati Public Schools provide this kind of evaluator training, as does the National Institute for Excellence in Teaching's Teacher Advancement Program (TAP) model.
- Collections of artifacts — lesson and curriculum-unit plans, assessments, and student work — that provide a basis for assessing such competencies as planning or alignment of curriculum to state standards. However, what constitutes a collection needs to be carefully limited so teachers don't feel compelled to develop a full portfolio.
- Multiple observations. Because teaching activities vary substantially over the day, week, and year, one observation per year is unlikely to provide a fair or representative sample of teaching practice. Although there is no magic number of observations that guarantees a good representation, three observations would seem a bare minimum. Having even two observations allows a check on how representative one observation actually is. In our own research, we found that four to five observations can give a high level of reliability.
- Multiple observers, including observers from outside the school. Research suggests that school administrator evaluations can't be free of leniency. Administrators need to maintain harmonious relationships and thus give marginal teachers the benefit of the doubt. Some administrators may never have seen really good teaching, and many high or middle school administrators may not have in-depth knowledge of all content areas. A content-knowledgeable observer from outside the school can add a more objective and informed perspective. The use of consulting teachers in Cincinnati and master teachers in Washington, D.C., are good examples of using observers from outside the school.

American Public University

You are 1 degree away from changing your world. Which 1 will it be?

79 affordable degrees of distinction – 100% online, including:

M.Ed., Teaching - Instructional Leadership (K-12)
M.Ed., Teaching - Special Education
M.Ed., Teaching - Curriculum & Instruction for Elementary Teachers
M.Ed., Teaching - Elementary Reading

Graduate level courses start monthly at only $300 to $325 per semester hour.

Start learning more at studyatAPU.com/education

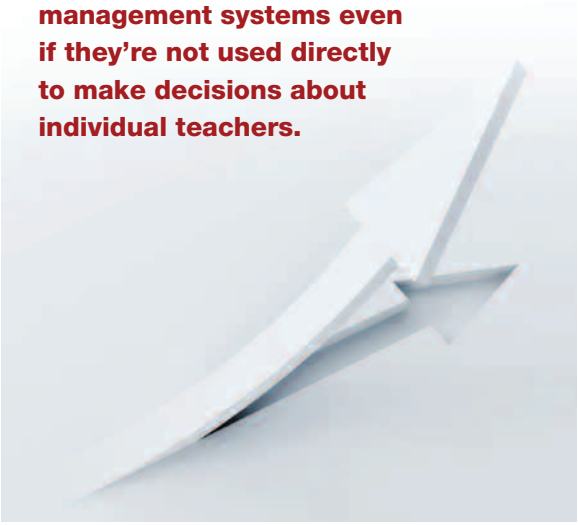
APU was recognized in 2009 and 2010 for excellence in online education by the prestigious Sloan Consortium.

APU American Public University

Text "APU" to 44144 for more info. *Message and data rates may apply.*

Many of these recommendations represent standard "best practice." There is now increasing evidence that ratings from carefully designed and implemented observation systems can be sufficiently reliable and valid for consequential uses.

Value-added and similar productivity measures can help improve human capital management systems even if they're not used directly to make decisions about individual teachers.



As well as promoting reliability and validity, evaluations must include features that help teachers learn from the results. Feedback should be specific and refer to the rubric or rating scale; it should enable teachers to understand why they received the scores they did. A trained person should be available to provide coaching and assistance to teachers who want to improve their performance, and other types of professional development should be readily available and linked to the competencies.

Policy makers also should consider the demands of these observation systems on administrators. Administrators' jobs may need to be redesigned to free up time. Help is needed from teacher leaders inside the school and evaluators from outside. The evaluation process should also be differentiated: New teachers and teachers who are struggling should receive a full dose of observation, feedback, and coaching every year, while experienced, tenured teachers might receive a full set of observations only every third year. Some of the video observation tools being developed by the Measuring Effective Teaching project may save time by allowing an "observation" to be uploaded to a web-based system to be scored by trained assessors.

Performance assessments or work samples. Performance assessments complement classroom observations by providing a more in-depth assessment of content knowledge, pedagogical content knowledge, use of formative assessment data, and differentiation of instruction. Teachers would demonstrate specific competencies in response to prompts or questions and would include such artifacts as unit or lesson plans, assignments, completed student work, and assessments.

Experienced and new teachers should be evaluated differently. A process adapted from the National Board for Professional Teaching Standards (NBPTS) assessments could be used for experienced teachers, while one based on the assessments developed by the Performance Assessment for California Teachers (PACT) Consortium could be used for new teachers. Scores on these measures have a positive relationship with measures of student learning (Wilson et al. 2007; Hakel, Koenig, and Elliott 2008). These assessments allow portfolios to be submitted electronically and reviewed during the summer by a group of content experts trained in scoring the items. This approach would mainly be used for such major decisions as tenure or movement on a career ladder.

The states should probably take the lead in developing and administering performance assessments. Not only are states more likely than districts to have the needed resources, they could use the result in multi-tier licensing systems. Also, the state determines the student content standards from which teacher content knowledge should be derived. Having states involved in educator assessments also reduces the likelihood of holding teachers in different districts to different standards.

Classroom walkthroughs. Even frequent formal observations can't provide enough information on typical instructional practice, especially about how key instructional strategies are routinely implemented. Classroom walkthroughs (brief, focused visits) are more efficient for this purpose. Walkthroughs get school leaders, instructional coaches, and mentors into classrooms frequently enough to see how teachers are developing and whether key instructional strategies are actually being implemented. If there are problems, school leaders can determine whether the problem is lack of skill (suggesting a professional development solution), lack of motivation (suggesting attention to setting goals and performance management), or context (perhaps teachers lack the time or resources). If data are collected systematically, they can be used to diagnose and to evaluate the response. Walkthroughs also provide opportunities to give more frequent formative feedback and encouragement to teachers and to recognize and reinforce good performance. Having school leaders, instructional coaches, or mentors frequently in classrooms looking for major elements of an instructional strat-

egy also sends the message that these things are important and that teachers are expected to implement them.

Many advocates of walkthroughs oppose using them to evaluate individual performance, but walkthroughs can improve summative teacher evaluations by providing a more representative sample of practice. It may be useful to designate some walkthroughs as nonevaluative (those done by instructional coaches and peers) and some as part of the evaluations (those done by administrators).

MEASURE PRODUCTIVITY AND PRACTICE

Over the past 20 years, "value added" statistical methods have made the idea of measuring teacher productivity credible. The basic idea is to estimate the contribution of a teacher to student learning by comparing the average achievement of a teacher's students to the level of achievement that would be expected for an average group of students with similar characteristics, including prior levels of achievement. The difference between the expected and actual level of achievement is an estimate of the "value-added" by a teacher. Most teacher value-added estimates are best interpreted as relative measures of productivity, since the expected achievement is generally based on an average.

While there are some major issues with value-added methods, they are the best productivity measures we have. Many states are moving ahead with them, and recent federal education policy (notably the Race to the Top competition) has promoted them. As districts and schools are increasingly held accountable for student achievement, some kind of outcome measure is needed that communicates the importance of getting results. At the individual teacher level, value-added methods are much more fair than measuring the average level of student attainment. But care must be taken in how value-added estimates are used.

Using value-added estimates of classroom productivity together with assessments of teaching practice is the best foundation for making management decisions about teachers, but they shouldn't just be averaged into one overall measure of teacher performance. Value-added and instructional practice measures represent two different constructs and have different measurement properties. The two scores can be *used* together, but *adding* them together would be like adding a person's weight and height.

How value-added and practice measures are used determines how they should be combined. A conjoint decision rule, such as requiring a minimum score on both measures, is a natural model to use for a tenure decision or movement on a career ladder. For example, a district might require a teacher to score "proficient" on the practice assessment by her third year and to have an average value-added score above a set level. For use in a termination decision about a tenured teacher, both the value-added and teaching performance scores would have to be low. However, setting cutoff points for the value-added measure will require substantial thought, because there is no natural cutoff point that represents acceptable performance.

Using the average value-added score as a cut point holds two problems. First, in most value-added systems, about half of the teachers will score below av-


• Administrative Program for Principals
• Superintendent Letter of Eligibility

California University of Pennsylvania, with more than 150 years of experience in higher education, offers two dynamic, 100% online programs created for you, the busy working professional looking to advance your career in education, but unable to fit traditional classroom study into your schedule.

All courses are taught completely online, in a cohort format, by experienced school administrators. You will form a professional relationship with members of your cohort as you begin and complete the program together. You can communicate with your professors and your cohort any time through online chats, bulletin boards and e-mail. Both part-time, completely online programs provide a personalized approach in meeting your needs, whether you are seeking a principal's certificate, a master's degree in education, the superintendent letter or simply a way to enhance your teaching and leadership skills.

To find out more about these two programs, or any of Cal U's Global Online programs, contact us at **866-595-6348**, e-mail us at **calugo@calu.edu** or visit **www.calu.edu/go.**

California University of Pennsylvania is accredited by the Commission on Higher Education of the Middle States Association of Colleges and Schools and the National Council for Accreditation of Teacher Education. Pennsylvania Department of Education-approved programs.

CALU
GLOBAL ONLINE

California University of Pennsylvania
Building Character. Building Careers.
www.calu.edu/go
A proud member of the Pennsylvania State System of Higher Education.

erage. This is probably too many to terminate or to deny tenure because it would be hard to replace this many. Second, value-added estimates, like nearly all performance measures, have a degree of error. Some of those who are just below average in the value-added distribution may really be above average, and some who show up as just above average may really be below. This error can be substantial, and it shows up as changes in teachers' rank in the value-added distribution from year to year (Goldhaber and Hanson 2008; McCaffrey, Sass, and Lockwood 2008).

Measuring teaching practice begins by translating a vision of effective instruction into a model that makes explicit what competent performance is.

Schools should use multiple years of value-added data for such decisions as tenure, pay raises, or termination, and the lower limit should be below the value-added average. For example, the minimum required for tenure might be set at the lower limit of the conventional 95% confidence interval. Decision makers could use a more narrow confidence interval if they're concerned more with avoiding tenuring teachers who later prove to be less effective (false positives) than with losing teachers who turn out to be better than predicted based on their initial performance (false negatives).

"*But I* want *to be left behind. It's too lonely without your friends.*"

Another option is to calibrate value-added in terms of the gains needed to move students to state proficiency standards. Administrators could use value-added estimates to develop expected trajectories for students and set required value-added levels high enough to maintain that trajectory. Districts would estimate how many teachers would be terminated and consider whether the additional terminations could be replaced without lowering the hiring bar.

Another approach for termination might be to use consistently low value-added scores as an initial signal that a teacher needs to be reviewed, no matter how high her or his practice assessment scores. Evaluators from outside a school would review the practice of teachers who received very low value-added scores. If the evaluator found that practice was below the "proficient" level, the teacher would have a year to improve practice. A teacher would have to improve to the "proficient" level to remain employed. Failing to improve the practice rating would lead to termination.

Combined value-added and practice measures would strengthen the argument for basing pay raises on knowledge and skill levels. Moving from one pay category to the next higher category would be based on evidence of both improved classroom practice and the effects on student learning.

For one-time pay bonuses, there is less need to combine value-added and teaching practice ratings. The stakes are lower because the reward is a one-time event. A natural model for this is a performance scorecard that simply reports the results of separate measures of performance (like a student's report card) and associates a bonus amount with achieving performance goals on each measure. This simple approach is easy to understand and allows teachers to be recognized for either practice or results, or both. The TAP model uses this approach.

OTHER USES FOR VALUE-ADDED MEASURES

Value-added and similar productivity measures can help improve human capital management systems even if they're not used directly to make decisions about individual teachers. Value-added teacher productivity estimates can be used to evaluate the effectiveness of such selection practices as web-based screeners. Even if value-added measures aren't used for tenure decisions, a district can use them to assess its tenure criteria by seeing whether the more productive teachers receive tenure. To reduce leniency in administrators' performance evaluation decisions, their ratings could be compared with value-added estimates. Showing an evaluator that value-added varies a lot while most teachers receive similar evaluation ratings might help raise awareness of leniency

and motivate reflection on how assessment decisions are made.

CONCLUSION

States and districts should develop practice measures by translating their visions of effective instruction into explicit competency models. These would include what teachers need to know and do to carry out state or district priorities. The model can then become the foundation for a set of practice measures that include observational rubrics for performance evaluation and management, performance assessments that would be part of tenure and pay systems, and walk-through tools for day-to-day performance management and for evaluating the implementation of instructional strategies. These tools can then be combined with measures of teaching productivity. Productivity measures should also be used to evaluate the quality of teaching practice measures and the effects of human capital management programs.

REFERENCES

Goldhaber, D., and M. Hansen. "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance in Making Tenure Decisions." *Policy Brief*. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, 2008. www.urban.org/publications/1001265.html.

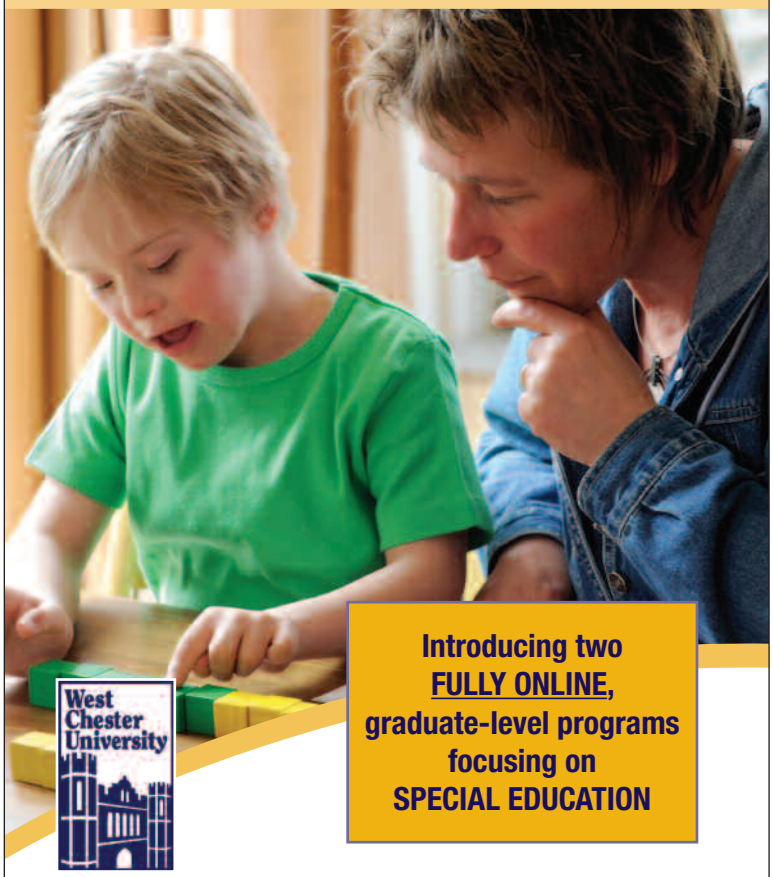
Hakel, Milton D., J.A. Koenig, and S.W. Elliott, eds. *Assessing Accomplished Teaching: Advanced-Level Certification Programs.* Washington, D.C.: National Academies Press, 2008.

McCaffrey, D.F., T.R. Sass, and J.R. Lockwood. *The Intertemporal Stability of Teacher Effect Estimates*. Nashville, Tenn.: National Center on Performance Incentives, 2008. www.performanceincentives.org/data/files/news/PapersNews/McCaffrey_et_al_2008.pdf.

Odden, Allan, and Marc Wallace. *How to Create World Class* Teacher Compensation. Minneapolis, Minn.: Freeload Press, 2008. www. textbookmedia.com/freeloadpress.

Wilson, Mark, P.J. Hallam, Ray Pecheone, and Peter Moss. "Using Student Achievement Test Scores as Evidence of External Validity for Indicators of Teacher Quality: Connecticut's Beginning Educator Support and Training Program." Manuscript, Stanford Center for Opportunity Policy in Education, 2007. http://edpolicy.stanford.edu/pages/pubs/pubs.html.

Unlock Their Your Potential



West Chester University

Introducing two FULLY ONLINE, graduate-level programs focusing on SPECIAL EDUCATION

M.Ed. in Special Education

and

POST-BACCALAUREATE CERTIFICATE in Universal Design for Learning and Assistive Technology

www.wcupa.edu/pdk

GRADUATE SPECIAL EDUCATION ONLINE PROGRAMS