Porter, J. M. & Jelinek, D. (2011). Evaluating Inter-rater Reliability of a National Assessment Model for Teacher Performance, *International Journal of Educational Policies*, *5*(2), 74-87.

ISSN: 1307-3842

Evaluating Inter-rater Reliability of a National Assessment Model for Teacher Performance

Jenna M. Porter[*]
California State University at Sacramento

David Jelinek[**]
California State University at Sacramento

Abstract

This study addresses the high stakes nature of teacher performance assessments and consequential outcomes of passing versus failing based on decisions of those who subjectively score them. Specifically, this study examines the inter-rater reliability of an emerging national model, the Performance Assessment for California Teachers (PACT). Current reports on the inter-rater reliability of PACT use *percent agreement* that *combines* exact and within 1 point agreement, but such measurements are problematic because adjacent scores of 1 point could be the difference between passing or failing. Multiple methods were used to examine the inter-rater reliability of PACT using 41 assessments (451 double scores) from an accredited institution in California. This study separated and examined the failing and passing groups, *in addition to* evaluating inter-rater reliability by combining them. Both percent agreement (exact and within 1 point) and Kappa (Cohen, 1960) were estimated to report the level of agreement among PACT raters for candidates who failed versus passed the assessment. Results indicate that inter-rater reliability ranged from poor to moderate, depending on whether a candidate passed or failed. A number of recommendations are proposed, including a model for more precise measurements of inter-rater reliability and improvements for training and calibration processes.

Keywords: Teacher performance assessment, Inter-rater reliability, PACT, High stakes assessments.

---

[*] Jenna Porter, Ph.D., California State University at Sacramento, Department of Teacher Education, 6000 J Street, Sacramento, CA 95819-6079, E-mail: jmporter@csus.edu

[**] David Jelinek, Ph.D., California State University at Sacramento, Department of Teacher Education, 6000 J Street, Sacramento, CA 95819-6079, E-mail: djelinek@csus.edu

Introduction

Just as classroom teachers are held accountable for student achievement, there is a growing movement to hold teacher education programs accountable for preparing quality teachers. Unfortunately, when the method for demonstrating proficiency is by testing every student using multiple choice assessments, very little critical thinking is required, and thus there is a gap between what is tested and what is actually taught (Wagner, 2008). The prominence of standardized tests influences instruction and emphasizes drilling for test preparation, which detracts from the amount of project based learning, inquiry, and higher level problem solving needed for critical thinking (Darling-Hammond, 2007). The direction of education reform in the United States that focuses on standardized testing is questionable, and limits the definition of success to the memorization of knowledge versus creativity, innovation, and emotional intelligence needed for the 21st century (Zhao, 2009). Assessments that require the *application* of knowledge may be a better method for evaluating what students understand.

The external accountability systems of teacher education programs has led to the development and use of standards-based performance assessments (Arends, 2006) in an effort to measure overall teaching competence. Traditionally, teachers have been required to pass subject matter competency and basic skills exams for licensure. However, measures of teaching competence that use records of coursework and paper-pencil exams are poor predictors of teaching effectiveness (Darling-Hammond, 2010). More comprehensive evaluations include teachers' actual performance in the classroom. One such measurement is a performance assessment. Performance assessments require demonstrating and applying essential knowledge to authentic tasks (Darling-Hammond, 2010; Mueller, 2011; Stiggins, 1987; Wiggins, 1993). For teachers, a good performance assessment would measure competencies related to the entire cycle of teaching to include planning, instruction, assessment, reflection, and subsequent application. Essentially, using performance assessments for evaluating teaching competency may be a better method of examining teachers' *reasoning* about the knowledge base they have learned.

However, performance assessments can be costly, time- consuming, and of little value if sound measurment methods are not implemented for their development and use (Ryan, 2006). Other problems associated with performance assessments may include unclear or incorrect performance criteria, inappropriate scoring methods, poor quality excersizes, and the training of evaluators (Stiggins, 1994). Some argue that performance-based assessments are considered to be more valid in general (Darling-Hammond, 2006; Darling-Hammond & Snyder, 2000) than other traditional measures of teaching ability, but they require raters to score subjectively. Thus, another relative concern that has not been sufficiently addressed is how the validity and reliability of these performance assessments are defined, measured, and reported. More research is needed, particularly since there is a widespread movement toward implementing a single standardized performance assessment to measure teacher competence.

Numerous teacher preparation programs across the country have authorized the use of standardized performance assessments for teacher licensure. Likewise, there is a

nationwide movement toward implementing and assessing a common set of standards for teaching competence. For example, the American Association of Colleges for Teacher Education (2011) has recently developed a national initiative, the Teacher Performance Assessment (TPA), which is aimed at gathering evidence of teaching competence and improving the consistency of licensure decision-making. Twenty four states including California, New York, Ohio, and Washington are represented in the consortium. The TPA assessment system includes the Performance Assessment for California Teachers (PACT), which is assumed to measure candidates' *application* of pedagogical strategies (Pecheone & Chung, 2006), and has been deemed a highly successful, valid, and reliable measure of teaching competence (American Association of Colleges for Teacher Education, 2011; Lombardi, 2011; Pecheone & Chung, 2007). However, using any single test for accountability purposes can be problematic when high stakes consequences exist for those who do not perform well, and some argue that PACT should not be used because its validity has not sufficiently been justified (Berlak, 2008; Campbell, 2007). Because an assessment cannot be deemed valid unless it is also a reliable measure of the identified construct, rater consistency should be highlighted when measuring and reporting the reliability of teacher performance assessments, which is a major objective of this study.

This study contributes to the advancement of education and public policy on teacher preparation by identifying alternative methods for analyzing inter-rater reliability of a high-stakes assessment such as PACT. Likewise, this study contributes to existing reports on PACT, and urges researchers to continue examining consequences of the assessments interpretation and use, so that teaching effectiveness can be fairly identified, particularly since national education policies are being made to include this test in their assessment system.

Emerging National Model: Performance Assessment for California Teachers (PACT)

Current reports on the inter-rater reliability of PACT and other performance assessments have included percent agreement (Educational Testing Service, 2003; Lombardi, 2011; Pecheone & Chung, 2007; Pecheone & Chung, 2006; Torgerson, Macy, Beare, & Tanner, 2009). However, inter-rater reliability has been reported using percent agreement that *combines* exact and within 1 point agreement. This is problematic because adjacent scores of 1 point could be the difference between passing or failing. Agreement within 1 rubric point is expected to occur and is still used to estimate levels of rater agreement. For PACT, this difference is acceptable for scores of 2, 3, and 4 (which are passing), but there is a major discrepancy when scores of 1 and 2 are counted as within 1 point, simply because 1 is failing and 2 is passing. Thus, it is imperative to evaluate and report exact agreement between raters in an effort to estimate inter-rater reliability more precisely.

Reports on rater consistency of PACT have also insufficiently addressed an important outcome variable: whether or not the teacher candidates pass or fail the assessment. Since candidates must pass PACT to earn their teaching credential, an examination of the inter-rater reliability of assessments that fail versus pass the test is

needed, which has not been reported in the literature. PACT is also a high stakes test, which is on the forefront of a nationwide movement to standardize the measurement of teacher competency, so its reliability and validity should be examined more closely, particularly relative to the consequential outcomes for those who fail.

Although performance assessments may be better measures of teaching competency over more traditional methods, mandating them in education policy for accountability purposes can be problematic. Implementing and requiring teacher candidates to demonstrate teaching competency by passing a performance assessment in their preparation programs is acceptable. However, the high stakes nature of the mandate and consequential outcomes of passing versus failing the assessment influences both teacher candidates and those who subjectively score them. Because an assessment cannot be valid unless it is also a reliable measure of a construct, an emphasis on inter-rater reliability of these performance assessments is needed to determine whether they should be adopted and used nationwide to grant or withhold teaching credentials.

The PACT Instrument

PACT is a teacher performance assessment that measures California teacher candidates' application of subject specific pedagogical knowledge. Many different versions of PACT exist to address the variety of credentials offered, such as bilingual and concurrent masters' programs, and to measure specific pedagogical knowledge in different subject areas. There are six different versions for multiple subject (elementary) and 18 versions for single subject (secondary) candidates, all of which are parallel and vary only by content area. All versions of PACT include a Teaching Event, where candidates must describe their classroom context (Context), develop a series of lessons in one content area (Planning), videotape themselves teaching a learning segment (Instruction), collect evidence and analyze student learning (Assessment), and reflect upon each of the tasks in a descriptive commentary (Reflection). Candidates' competence in addressing the Academic Language task is also measured and is woven throughout all tasks in the Teaching Event. The only difference among PACT's multiple versions is an emphasis on subject-specific pedagogy. For example, the Planning task in Elementary Math requires teachers to reflect upon how they plan learning activities that build on each other to support students' development of conceptual understanding, computational/procedural fluency, and mathematical reasoning skills. On the other hand, the Planning task in Elementary Literacy asks teachers to reflect upon how they plan learning tasks that build on each other to develop students' abilities to comprehend and/or compose text (PACT Consortium, 2008). This study analyzed data from multiple subject Elementary Mathematics and eight different single subject versions including Art, English, Math, Music, Physical Education, Science, Social Science, and World Languages.

PACT Scoring and Calibration

A task-based scoring model is used to score the Teaching Event. Each task addresses specific pedagogical knowledge and is scored using a 4-point rubric scale. Raters

evaluate candidates' competence based on multiple sources of data from narrative commentaries, videotapes, lesson plans, and work samples. In order to pass the Teaching Event, candidates must pass all five tasks. A score of 1 is considered failing and indicates that the candidate has not met the teaching standard. A 2 is passing and means the candidate minimally meets the standard. Candidates scoring 3 or 4 illustrate advanced levels of standards. In order to pass the Teaching Event, candidates must earn a majority of passing scores (2 or above) within the task and have no more than two failing scores (1) across tasks. This means that a candidate can pass the entire Teaching Event but still have one or two failing scores (out of 11 possible).

Rater training and calibration protocol are purportedly standardized to maximize validity within programs and across all universities using PACT (Pecheone & Chung, 2006). The PACT Consortium (2009) developed a document titled, "Thinking Behind the Rubrics" to help raters apply the evidence collected from the Teaching Event to the rubric levels. It discusses big ideas for each rubric and identifies key differences between adjacent rubric scores. "Thinking Behind the Rubrics" is supposedly used for training raters, and is encouraged for use during scoring. Subject-specific Teaching Events are also selected as *benchmarks* for training and calibration.

Raters are calibrated based on criteria that their scores must result in the same pass/fail decision, and cannot include any scores that are 2 points away from the predetermined *benchmark* score. Raters are required to calibrate once per academic year with other subject-specific raters within or across universities. However, due to different lengths of credential programs and other factors related to timing and resources, scoring sessions might occur during various times throughout the academic year. The time gap between calibration and scoring could be as short as one day or longer than six months, depending on how teacher education programs interpret protocol.

Inter-rater Reliability of PACT

Inter-rater reliability is one type of internal consistency measure that differs from scale reliability in that it evaluates the level of agreement among raters versus the reliability of the assessment itself. Inter-rater reliability can be operationalized in different ways to include the agreement with the "official" score and/or agreement with other raters. Calibrating raters may be one method for improving levels of agreement among them, but the calibration standard for PACT is set in a way that covers nearly the entire scale, which will be addressed further in the discussion section.

All accredited institutions in California are required to double score at least 15 percent of their TPAs to check inter-rater reliability. One potential issue related to this is that failing and passing Teaching Events have been combined in reporting the double scores. The California Commission on Teacher Credentialing does not require the two groups to be separated, which is problematic since there is a systematic difference between them: one group failed the assessment, requiring an automatic double score, and the other passed and was randomly selected for rescoring. Reports on PACT to date have combined these two groups into one category for evaluating inter-rater reliability but it is important to analyze these groups separately to learn more about levels of

agreement between raters of failing Teaching Events versus those that pass. Furthermore, combining the two categories may lead to the misrepresentation of PACT's inter-rater reliability because failing Teaching Events are overrepresented, which will be addressed further in the discussion section.

Percent Agreement

Pecheone & Chung (2007) evaluated score consistency of PACT data from the pilot years by computing inter-rater agreement percentages. The first pilot year yielded 395 Teaching Events, 41 percent of which were double scored to evaluate inter-rater reliability. Score matches were exact or within 1 point 91 percent of the time. However, only 57 percent of the scores yielded exact matches, while an additional 34 percent matched within 1 point. What is worrisome is that inter-rater reliability reports of PACT have reported very high levels of agreement (91 percent) amongst raters but combine exact and within 1 rubric point matches into a single category, and do not further examine the within 1 point differences. This is problematic as a measure of inter-rater reliability since a 1 point difference in PACT scores could yield either passing or failing results. As an example, the first rater could give a candidate a failing score of 1 on one rubric while the second rater gives a passing score of 2. So combining exact and within 1 point agreement percentages is misleading since a discrepancy could exist between a passing or failing decision, which indicates poor, not strong levels of agreement, as suggested by the single combined percent.

Kappa

One method that can be used for estimating inter-rater reliability is Kappa (Cohen, 1960). Kappa is the proportion of observed agreement minus chance agreement divided by one minus chance agreement, as illustrated below:

$$(\kappa = \frac{P_o - P_e}{1 - P_e})$$

Chance agreement is estimated by the number of agreements expected assuming ratings from different raters are completely random. The statistic also assumes that raters are independent observers. As the number of possible scores for observed ratings decrease, chance agreement increases. The Kappa coefficient ranges from -1.00 (perfect disagreement) to 1.00 (perfect agreement), and can be interpreted to indicate a percentage of agreement that is accounted for from raters above what is expected by chance. A Kappa of 0 would mean that raters have a random level of agreement or disagreement, such that there is no relationship between their ratings. Some consider Kappa an improvement over percent agreement (Bryington, Palmer, and Watkins, 2002). One benefit is that it can be compared across different conditions since it is corrected for chance (Ciminero, Calhoun, & Adams, 1986). This advantage of Kappa would be particularly useful in examining PACT scores since credential programs vary across universities. Raters can also vary tremendously in terms of their experiences and subject area expertise, especially considering the multiple versions of PACT that must be scored annually.

There is some disagreement, however, over how to interpret "fair agreement" with the Kappa statistic (see Table 1). Fleiss (1981) suggests a very strict guideline for

fair agreement, such that Kappa must be greater than .41 to yield any type of acceptable agreement. Anything less would be considered poor or weak agreement. Landis and Koch (1977) suggest criteria that have the most descriptive range, while Altman (1991) suggests anything below .20 as poor agreement. They all agree, though, that the Kappa coefficient should be greater than .61 to interpret good to substantial agreement.

Table 1

Agreement Levels of Kappa

	<0.0	0.0-.20	.21-.40	.41-.60	.61-.80	.81-1.0
Altman, D. (1991)	Poor	Poor	Fair	Moderate	Good	Very Good
Landis & Koch (1977)	No agreement	Slight	Fair	Moderate	Substantial	Almost perfect
Fleiss (1981)	Poor	Poor	Poor	Fair	Good	Excellent

Methods and Analyses

Multiple methods were used to examine the inter-rater reliability of PACT using 41 assessments (451 double scores) from one accredited institution in California. This study separated and examined the failing and passing groups, *in addition to* evaluating inter-rater reliability by combining them. Both percent agreement (exact and within 1 point) and Kappa (Cohen, 1960) were estimated to report the level of agreement among PACT raters for candidates who failed versus passed the assessment.

Participants and Materials

A total of 181 teacher candidates at a California State University participated in the study. There were 87 candidates seeking single subject credentials in a variety of content areas including History, Science, and Art. Remaining candidates (*n*=94) were enrolled in multiple subject programs to teach in Kindergarten through sixth grade. Approximately 25 university faculty members scored the PACT Teaching Events. All raters were trained and calibrated in the content areas they were scoring. Scores from the Performance Assessment for California Teachers **(**PACT) were the sole source of data for this study.

Procedures

Double scoring. Following initial scoring of all 181 Teaching Events, a total of 23 percent were double scored. An average 11 percent (*n*=20) of candidates had failing Teaching Events and needed to be double scored. Another 12 percent (*n*=21) of Teaching Events were randomly selected for double scoring. The double scoring of both failing and passing Teaching Events occurred subsequent to initial scoring. The first round of scoring was organized by the PACT coordinator, who assigned Teaching Events to match rater expertise. Once failing and passing Teaching Events were

identified, a second round of scoring occurred (within approximately one week of each other). Thus, second round double raters were unaware whether the Teaching Events they were scoring were failing or passing.

Analysis. Multiple methods were used to calculate inter-rater reliability. Percent agreement was estimated two ways. First, exact matches and agreement within 1 point were combined for comparison to existing studies. But because combining agreement within 1 point potentially yields one passing and one failing score, exact agreement was also calculated as a more precise measure of rater consistency, and to identify the number of discrepancies between passing versus failing scores. Kappa was also estimated as an additional measure of inter-rater reliability to account for chance agreement. All results were also categorized by failing versus passing Teaching Events to examine any potential differences between them.

Results

Overall, agreement levels of PACT raters ranged from poor to moderate, depending on whether a candidate passed or failed. Two major findings emerged: (a) combining exact and within 1 point agreement yielded discrepancies between passing and failing, and (b) inter-rater reliability differed in terms of whether a candidate failed or passed.

Percent Agreement

Percent agreement for all double scores was calculated for the overall sample and by failing versus passing Teaching Events (see Table 2). The combined percentage agreements (exact and within 1 rubric point) indicated strong rater agreement. Reported alone, it might appear that the inter-rater reliability of PACT is near perfect. However, exact matches and within 1 point agreement become more descriptive when evaluated separately. Even though the combined percent agreements were near perfect, approximately one third of the score pairs were within 1 point.

Table 2

Percent Agreement of PACT for Failing, Passing, and Overall Categories

	Exact	Within 1 point	Combined
Failing	61%	33%	94%
Passing	71%	26%	97%
Overall	66%	30%	96%

The category within 1 point was examined further for any discrepant score pairs (one rater assigned a failing score and a different rater assigned a passing one). Agreement within 1 rubric point is expected to occur and is still used to estimate levels of rater agreement. For PACT, this difference is acceptable for scores of 2, 3, and 4 (which are passing), but there is a major discrepancy when scores of 1 and 2 are counted as within 1 point, simply because 1 is failing and 2 is passing. A percentage of the within 1 point category yielded discrepant score pairs (see Table 3).

Table 3

Rater Pass/Fail Discrepancy Percentages

	N cases	Discrepant score pairs
Failing	209	34%
Passing	242	6%
Overall	451	20%

Overall, 20 percent of the double scored Teaching Events yielded pair matches of 1 and 2, which was determined discrepant. Teaching Events that passed yielded much less discrepancy, whereas raters of the failing Teaching Events tended to disagree more on whether a candidate should pass or fail the rubric item. This may be expected, since the groups were separated and the failing Teaching Events were originally assigned more scores of 1, potentially leading to higher discrepancies. However, it remains problematic because 20 percent of the double scores were discrepant in passing versus failing for the overall sample. Thus, combining all of the within 1 point matches may not be a sound method for estimating percent agreement since the slight disagreement of 1 point could be the difference between a passing or failing rubric. One area of concern regarding this combination is PACT's 4-point scale. The range of passing scores is higher than failing, since a score of 1 is the only failing possibility. Agreement within 1 point is half of the scale, and the calibration standard is set in a way that a range of three different scores is acceptable for a single rubric item. In general, when Teaching Events were separated into failing and passing categories, there was less exact agreement for raters of failing Teaching Events than for raters of the passing group.

Kappa

Kappa coefficients were consistent with exact agreement results. The Kappa coefficient for the overall sample yielded fair results (see Table 4). However, when passing and failing assessments were separated, raters of failing Teaching Events yielded poor agreement, whereas the passing group yielded moderate agreement.

Table 4

Kappa Estimates for PACT

	Kappa	P	Interpretation of Kappa
Failing	0.18	0.00	Poor
Passing	0.41	0.00	Moderate
Overall	0.35	0.00	Fair

According to several researchers, good agreement coefficients should be greater than 0.60. (Altman, 1991; Fleiss, 1981; Landis and Koch, 1977), yet findings suggest moderate agreement at best.

Conclusion

Findings of the study suggest that the inter-rater reliability and consequential validity of PACT should be evaluated further, since results illustrated major discrepancies between what raters considered passing versus failing. Moreover, inter-rater reliability of performance assessments such as PACT should be estimated and reported using more precise methods that do not combine exact and within 1 point agreement, *and* separate results into pass and fail categories.

Because there was less agreement between raters of failing assessments, there is a need for further examination of the cut score (or passing standard for PACT) in relation to rater agreement. Essentially, the evaluation of the cut score is a key issue in validating the decisions made from it (Kane, 2006). Assuming that PACT is a reliable and valid measure of teaching effectiveness, failing the assessment would indicate that a candidate had not met the level of competence deemed necessary to become a teacher. But since there is such rater disagreement on whether to pass or fail candidates, the passing standard and procedures for identifying "incompetent" candidates is questioned and needs further examination. Likewise, future studies on the inter-rater reliability of PACT should report rater consistency by separating passing versus failing assessments.

The unitary concept of validity (Messick, 1989; Moss, 1998) is generally accepted but researchers continue to propose new types (Suen & French, 2003). One element of validity that particularly relates to PACT is consequential validity. Very few reports on the consequential validity of tests, including PACT, exist in the educational measurement literature (Reckase, 2005), and it is not emphasized sufficiently in test development or use (Barnett, Macmann, & Lentz, 2003). Examining the consequential validity of any assessment used for high stakes is necessary. Consequential validity can be defined as one element of validity that addresses the social issues and consequences of test interpretation and use (Cronbach, 1982; Kane, 2006; Messick, 1989; Moss, 1998). Consider the following scenario: A subgroup of candidates fail the PACT and consequently do not earn their teaching credentials. This does not necessarily mean that PACT is invalid. However, if it turned out that PACT measured something different for this failing subgroup than for the passing group, the construct validity would be jeopardized, therefore influencing its consequential validity as well. Essentially, PACT has to measure what it says it is measuring for all subgroups, and because pilot studies relative to the bias and fairness review indicated some significant differences between gender and the socioeconomics of schools, it is not sufficient to say that the validity of PACT is acceptable. Thus, not enough research has been conducted in terms of validity to suggest that the assessment alone identifies competent teacher candidates.

Although more comprehensive evaluations of teachers' performance in the classroom using performance assessments may be better indicators of teaching effectiveness than traditional paper-pencil assessment measurements, the high-stakes consequences of such assessment measures demand careful scrutiny of reliability

between raters. Results from this study suggest that current methods used to calculate the inter-rater reliability of the PACT Teaching Event produce a very high level of rater reliability. But using more precise measurements highlighted major shortcomings in previous methods, especially when discrepancies between what raters considered passing versus failing are examined.

As this study illustrates, the use of Kappa measurements to estimate inter-rater reliability, plus an examination of the discrepancies between what raters considered passing versus failing, produce a more accurate picture of inter-rater reliability. Both percent agreement (exact and within 1 point) and Kappa were estimated to report the level of agreement among PACT raters for candidates who failed versus passed the assessment. Combined percent agreement (of exact and within 1 point agreement) reported alone provides a misleading picture because it leads to the conclusion that the raters are in near perfect agreement, but in actual fact, this study bears out by using more precise measurements, that the overall agreement levels of PACT raters ranged from poor to moderate. One major conclusion of this study, therefore, is that exact matches and within 1 point agreement should be evaluated separately.

Another major conclusion is that the training and calibration of raters needs attention. Raters of this study were calibrated yet yielded a poor level of inter-rater reliability for failing Teaching Events. This points to the need for further examination of the standardized PACT training and calibration protocols and how they are being interpreted and implemented.

Finally, the issue of consequential validity of PACT should be further addressed by examining whether the failing group of candidates differed systematically than those who passed. It is the responsibility of the test developers *and* users to collect evidence of the consequences of the test (Kane, 2006; Nichols & Williams, 2009). The high stakes and summative nature of PACT may influence teacher performance on the assessment, but this issue has rarely been examined in educational research. It is important to consider how an assessment's purpose influences teacher candidates' experiences with the assessment in terms of learning outcomes (Chung, 2008). Although some validity analyses were conducted on PACT (Pecheone & Chung, 2006), there were questionable results regarding differences between particular subgroups of candidates. Thus, current research on PACT does not adequately address validity such that claims can be made for its suitable use to grant or withhold teaching credentials.

References

Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hal

American Association of Colleges for Teacher Education (2011). *Teacher performance assessment consortium.* Retrieved May 30, 2011 from http://aacte.org/index. php?/Programs/ Teacher-Performance-Assessment-Consortium-TPAC/teacher-performance-assessment-consortium.html

Arends, R. (2006). Performance assessment in perspective: History, opportunities, and challenges. In S. Castle & B. Shaklee (Eds.), *Assessing teacher*

performance: Performance-based assessment in teacher education (pp. 3-22). Lanham, MD: Rowman & Littlefield Education.

Barnett, D. W., Lentz, F. E., Jr., & Macmann, G. M. (2000). Psychometric qualities of professional practice. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral Assessment* (2nd ed., pp. 355-386). NY: Guilford.

Berlak, A. (2008). *Assessing PACT,* A state-mandated teacher performance assessment for identifying highly qualified teachers. Retrieved May 30, 2011 from http:// sites.google.com/site/assessingpact/Home

Bryington, A, Palmer, D., Watkins, M. (2002). The estimation of interobserver agreement in behavioral assessment. *The Behavior Analyst Today*, 3, 323-328).

Campbell, D. (2007). *Problems with legislation on teacher performance assessment (TPA), PACT, and some suggestions*. Unpublished manuscript.

Chung, R. (2008). Beyond Assessment: Performance assessments in teacher education. *Teacher Education Quarterly*, Winter, 7-28.

Ciminero, A. R., Calhoun, K. S., & Adams, H. E. (Eds.). (1986). *Handbook of behavioral assessment* (2nd ed.). New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20 (1), 37–46.

Cronbach, L. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Darling-Hammond, L. & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*, 523-545.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education,* Vol. 57 (2), 120-138.

Darling-Hammond, L. (2007, May 21). Evaluating "No Child Left Behind." *The Nation.*

Darling-Hammond, L. (2010). Evaluating Teacher Effectiveness: How teacher performance assessments can measure and improve teaching. Retrieved from http://www.americanprogress.org/issues/2010/10/pdf/teacher_effectiveness.pdf

Educational Testing Services. (2003). *Scoring analysis for the field review of the California teaching performance assessment*. Retrieved March 30, 2009 from http://www.ets.org/Media/About_ETS/pdf/scoring. Pdf

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. NY: Wiley.

Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th edition, pp. 17-64). Washington, DC: American Council on Education/Praeger.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 33, 159-174.

Lombardi, J. (2011). *Guide to Performance Assessment for California Teachers (PACT*). Boston, MA: Allyn & Bacon/Merrill.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.13-103). New York: American Council on Education and Macmillan.

Moss, P. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice,* 17(2), 6-12.

Mueller, J. (2011). Authentic Assessment Toolbox. Retrieved August 9, 2011 from http://jfmueller.faculty. noctrl.edu/toolbox/whatisit.htm

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence:Roles and responsibilities. *Educational Measurement: Issues and Practice*, *28*(1), 3-9.

PACT Consortium. (2008). *Teaching Event Handbooks.* Retrieved May 15, 2009 from http://www.pacttpa.org

PACT Consortium. (2009). *Thinking Behind the Rubrics*. Retrieved October 9, 2009 from http://www.pacttpa.org

Pecheone, R., & Chung, R. (2006). Evidence in teacher education. *Journal of Teacher Education,* 57(1), 22-36.

Pecheone, R., & Chung, R. (2007). *Performance Assessment for California Teachers (PACT): Summary of validity and reliability studies for the 2003- 04 pilot year*. (Technical Report: PACT Consortium, 2007). Retrieved January 28, 2009 from http://www.pacttpa.org /_files/Publications_and_Presentations /PACT_Technical_Report_March07.pdf

Reckase, M. (2005). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice,* 17(2), 13-16.

Ryan, T. (2006). Performance Assessment: Critics, Criticism, and Controversy. *International Journal of Testing,* 6(1), 97-104.

Stiggins, R. J. (1987). The design and development of performance assessments. *Educationa Measurement: Issues and Practice, 6*, 33-42.

Stiggins, R.J. (1994). *Student-centered classroom assessment.* New York: Macmillan Publishing Company.

Suen, H. K., & French, J. L. (2003). A history of the development of psychological and educational testing. In C.R. Reynolds & R. Kamphaus (Eds.) *Handbook of Psychological and Educational Assessment of Children, 2nd ed.: Intelligence, aptitude, and achievement (pp. 3-23).* New York: Guilford.

Torgerson, C., Macy, S., Beare, P., Tanner, D. (2009). Fresno assessment of student teachers: A teacher performance assessment that informs practice. *The Free Library*. (2009). Retrieved January 30, 2010 from http://www.thefree library.com/Fresno assessment of student teachers: a teacher performance

Wagner, T. (2008). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need – and what we can do about it.* New York: Basic Books.

Wiggins, G. P. (1993). *Assessing student performance*. San Francisco: Jossey-Bass Publishers.

Zhao, Y. (2009). Catching up or leading the way: American education in the age of globalization. Alexandria, VA: ASCD.

About the Authors

Dr. Jenna Porter has served as a public school teacher, and currently works as a lecturer in the Teacher Credentials Department at California State University, Sacramento. She teaches pedagogy, urban education, science methods, critical thinking, and school and community courses. She also coordinates student teaching placements for the single subject credential program, in addition to supervising teacher candidates in the field. Her research interests include reflective practices in teacher education, lesson study, socio-cultural perspectives on learning and school community partnerships.

Dr. David Jelinek has been a K-12 teacher, principal, dean and professor. He teaches educational psychology, methods and assessment, professional development, pedagogical, sociological and research methods courses. He is the ePortfolio Manager for the College of Education's assessment system, which involves overseeing the cloud-based system that collects, assesses and reports student formative and summative assessments. He is also the College of Education PACT Co-coordinator, in which he oversees the electronic collection and assessment of teaching credential candidate's state-mandated performance assessment. Dr. Jelinek has been the principal investigator for several grant projects, including an NSF grant, and technology-related programs. Most recently, he completed a two-year project at Chaminade University of Honolulu to develop and implement an online graduate program for their education division. He is currently under contract with Pearson Publishing to develop an interactive textbook that will integrate cloud-based technology into an electronic textbook.

International Journal of Educational Policies

ISSN: 1307-3842

http://ijep.icpres.org
http://ojs.ijep.info