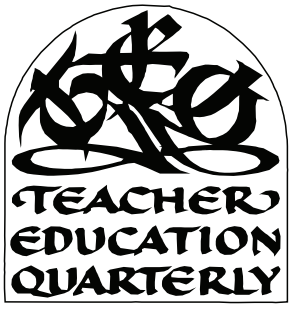

Critics and Critical Analysis:
Lessons from 19,000 P-12 Students in Candidates' Classrooms

By Jacqueline Waggoner, James B. Carroll, Hillary Merk, & Bruce N. Weitzel

Ever since the 2000 revision to the National Council for the Accreditation of Teacher Education's (NCATE; 2013b) standards, schools of education have searched for the most productive ways to measure candidates' impact on student learning. This has been no easy task (Hamel & Merz, 2005). Although candidates are often in student teaching experiences for the better part of a year, the ability to measure candidates' impact on student learning is mitigated by the degree to which the cooperating teacher or university supervisor assists the candidates, the length and continuity of instruction that the candidates provide, and the type and number of assessments the candidates may use. These problems are exacerbated by the difficulties in gathering student learning data that are comparable across candidates, programs, and even schools of education.

NCATE realized these problems but did not remove the expectation for measuring candidates' impact on student learning. It suggested a cluster of candidate activities that might provide the required evidence. A candidate (a) undertakes a diagnosis (a pretest) or P-12 student learning in some area he or she will teach; (b) plans an

Jacqueline Waggoner is an associate professor, James B. Carroll is a professor, Hillary Merk is an associate professor, and Bruce N. Weitzel is an associate professor and associate dean, all with the College of Education at the University of Portland, Portland, Oregon. waggoner@up.edu, carroll@up.edu, merk@up.edu, & weitzel@up.edu

appropriate sequence of instruction to advance P-12 student learning and teaches in ways that engage P-12 students, who bring differing background knowledge and learning needs; (c) conducts some concluding assessment (or posttest) that documents that student learning has occurred, or has not; and (d) reflects on changes in teaching that might have improved the results (NCATE, 2013a, Assessment 5).

The merger of the Teacher Education Accreditation Council with NCATE into the Council for the Accreditation of Education Preparation (CAEP) included standards that continue the focus on measuring candidates' impact on student learning (Council for the Accreditation of Educator Preparation, 2013):

> *Impact on P–12 student learning*
> 4.1 The provider documents, using value-added measures where available, other state-supported P-12 impact measures, and any other measures constructed by the provider, that program completers contribute to an expected level of P-12 student growth.

No consistent strategies for gathering student learning data have appeared, even though these accreditation standards have existed for some time. The highly promoted Education Teacher Performance Assessment (edTPA) focuses on using student summative assessment for the whole class and for three students in candidates' classrooms to understand the impact of instruction (American Association for Colleges of Teacher Education, 2013). Those data can vary in type and richness, reducing the comparability across candidates or programs; nor are these data measures of student learning gains per se.

Value-added measures (VAM) have been added into the accreditation language over recent years. These measures include the analysis of standardized assessments of student learning designed to address differences in classrooms and students to have equitable comparisons of teacher impact. In most cases, VAM would not be possible for candidates to use because of the small amount of time that candidates student-teach by themselves with the curriculum for which the standardized assessment is designed. Louisiana, among other states (Knight et al., 2012), has addressed this problem by applying VAM assessments to the P-12 students of their candidates who have been tracked into their first years of teaching. Regardless of the efficacy of this approach, it does not address the difficulties that schools of education have tracking an individual candidate's impact on student learning while student teaching and tracking the impact of the program on the candidate's progress over the time the candidate is in the unit's program.

After an NCATE review in 2007, we began an effort to have candidates report results on P-12 assessments aligned with units of instruction. After a lengthy development process, for 5 years, we have systematically gathered data on learning gains for each P-12 student in every teacher candidate's classroom. The result is a database of demographic data and learning gain scores for 19,334 P-12 students. These data provide a rich resource for understanding the progress of our candidates and the

impact of our programs. This study examines the impact on program improvement of systematically gathering P-12 student learning data over a 5-year period.

Methods

This study was completed over 5 academic years (2008-2009 to 2012-2013) in a teacher preparation program in northwest Oregon. Data were gathered from two student teaching experiences of teacher candidates during the fall and spring semesters of their practicum experience. The candidates were completing either a 10-month master of arts in teaching (MAT) program or a 4-year, undergraduate licensure program (see Table 1). Practica in these programs occurred in both private and public schools in Oregon and Washington. Oregon divides teacher certification into four levels: early childhood, elementary school, middle school, and high school. Candidates receive authorization at two levels of certification. Placements of candidates in this study represented all four of these levels.

In Oregon, candidates are required to prepare and teach a unit of instruction during each of the student teaching experiences. The design of these units of instruction followed the teacher work sample methodology. Candidates gathered data on the context of the school in which they were teaching, wrote goals for the unit of instruction based on Oregon State curriculum standards, designed and delivered instructional activities for the unit, prepared and administered preassessments and summative assessments of student learning, were video recorded teaching a lesson, and wrote prompted reflections on the process. The work sample was prepared as an artifact of the student teaching experience, and the evaluation of the work sample was a major component of measuring candidate readiness for teaching.

In 1997, the State of Oregon rewrote the administrative rules governing teacher licensure. A description of the specific requirements for the work sample was included in that revision. As part of the assessment requirements, candidates were instructed to gather "data on learning gains resulting from instruction, analyzed for each student, and summarized in relation to students' level of knowledge prior to instruction" (Oregon Teacher Standards and Practices Commission, 2013, OAR

Table 1
Candidates by Program and Semester

	2008-2009		2009-2010		2010-2011		2011-2012		2012-2013	
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Undergrad.	36	38	43	42	34	35	20	28	25	30
MAT	36	40	66	61	46	40	29	36	26	34
Total	72	78	109	103	80	75	49	64	51	64

Note. MAT = master of arts in teaching program.

584-017-1030). The candidates in this study were specifically asked to measure levels of student knowledge at the beginning of the unit of instruction with a preassessment and then to use a matched summative assessment when the unit concluded.

At the end of each of the student teaching experiences, candidates filled out a preformatted Excel spreadsheet that included information for each of the students in their classrooms on gender, ethnicity, identified learning needs, the preassessment score, and the summative assessment score. Learning needs were coded following No Child Left Behind guidelines as English language learner, talented and gifted, special education, and those students on 504 Learning Plans. To compare progress that these students were making with the progress of students who did not have identified learning needs, a fifth coding category of no identified need was used for all other students. The Oregon licensure authorization level was identified for each candidate classroom experience. Candidates produced work samples in placements at two levels of authorization within their programs (early childhood, elementary school, middle school, or high school). The fall experience was in their second preference level of authorization, and the spring experience was in their preferred level. Additionally, the content area of the unit of instruction (analysis was limited to the core areas of language arts, social studies, math, and science) was identified. After candidates electronically submitted their spreadsheets to a records clerk, we identified the socioeconomic status (SES) of the school in which the candidate had been working. Oregon rank orders all schools by SES using a computation based on four measures of economic status of families in the school, and quartile rankings were developed from this list. For schools that were not included on this list (private schools and candidate placements in southwest Washington), median family income for the school community was compared to ranked Oregon schools to assign a comparable SES quartile.

The candidate listed on the spreadsheet the maximum possible score for each assessment. Student scores were then translated into percentage correct scores. The summative assessment scores were used as the best indicator of whether students had met unit goals. Learning gains were computed as the difference between the preassessment and the summative assessment scores. Data from each of the submitted Excel spreadsheets were reformatted and transferred to a single worksheet. The final database included data on gender, ethnicity, identified learning needs, content area of instruction, school SES, authorization level, postassessments as percentage correct, learning gains as percentage increase, the program in which the candidate was enrolled, and the semester in which the unit of instruction was completed.

Summative assessment scores and learning gains (differences between pretest and posttest percentage correct scores) were used as dependent variables in analyses of variance of each of the demographic variables (á=.05). These analyses were repeated for both fall and spring data. Bonferroni post hoc analysis was used because of the repeated analysis of the same dependent variable data (Castaiieda, Levin, & Dunham, 1993).

Discussion of Effect Size (Cohen's d)

Effect size was determined using pooled standard deviations divided into mean differences of statistically significant ANOVA comparisons—Cohen's *d* (Cohen, 1988). The identification of statistical significance in large samples is problematic. Kish (1959) stated,

> In small samples significant, that is, meaningful, results may fail to appear *statistically significant*. But, if the sample is large enough the most insignificant relationships will appear statistically significant. . . . The word significance should be attached to another question, a substantive question, Is the relationship here worth explaining? (p. 336)

In this study, sample sizes are inordinately large for studies of educational phenomena. We expected the results to demonstrate statistical significance in almost every comparison but wanted to focus on those comparisons that had effect sizes large enough to warrant further examination. Our hope was that evidence would show no statistical differences, indicating candidates were addressing the needs of all P-12 students equitably. This seemed unlikely to happen because of the very large number of P-12 students from whom we had data. We adopted Cohen's (1988) view of effect sizes of d=.2 as representing small effect sizes. An effect size greater than .2 potentially represents real differences that could indicate inconsistencies in how candidates impacted student learning.

Disposition of Results for Continuous Program Improvement

Results of the yearly analysis of data were presented to the School of Education Assessment Committee, and recommendations for program improvement based on the data were forwarded to the full faculty of the School of Education for approval and implementation.

Results

Preassessment and summative assessment scores were reported for 19,334 P-12 students over a 5-year period in 720 classrooms led by 360 teacher candidates. Scores were converted to percentage correct to standardize results across candidate experiences. Learning gain scores representing the difference between preassessment and summative assessment percentage correct scores were computed. There were 19,208 useable learning gain scores with a grand mean of 35.17 and a standard deviation of 25.85. The percentage learning gain scores ranged from -66.67 to 100.00. ANOVA analysis of percentage learning gain scores was completed for each of the independent variables: program type, program semester, level of endorsement, school SES, gender, ethnicity, learning needs, and content area.

Five-Year Data Summary

Mean percentage learning gains for P-12 students by gender was 37.64 for girls and 36.59 for boys (see Table 2). Scores were statistically significantly different (p=.005) with a very small effect size of .04.

Mean percentage learning gains by ethnicity (see Table 3) were statistically significantly different for Black and Hispanic students with White (non-Hispanic) students (p<.001) and Asian students (p=.013 when compared with Black and p=.041 when compared with Hispanic). Effect sizes did not exceed .13. The percentage of non-White (including Hispanic) students in this study was 38.2%. The 2010 Portland-area non-White (including Hispanic) population was reported to be 27.8% (U.S. Census Bureau, 2014).

Mean percentage learning gains by learning needs (see Table 4) showed that English language learners and students identified in special education had statistically significantly lower learning gain scores than students identified as talented and gifted and students for whom no learning need was identified (p<.001). Effect sizes were .15 for talented and gifted comparisons and .18 for comparisons with students with no identified learning need.

Table 2
Mean Percentage Learning Gain for P-12 Students by Gender

	n	*M*	*SD*
Girls	9,014	37.64	25.40
Boys	8,786	36.59	25.22

Note. p=.005; Cohen's d=.04.

Table 3
Mean Percentage Learning Gain for P-12 Students by Ethnicity

				p (Cohen's *d*)	
	n	*M*	*SD*	Black	Hispanic
Black	1,387	34.69	24.65		
White (non-Hispanic)	10,892	38.02	25.13\	<.001 (.13)	<.001 (.11)
Hispanic	2,293	35.27	25.79		
Asian	1,395	38.02	26.57	.013 (.13)	.041 (.10)
American Indian/ Pacific Islander	261	36.76	27.61		
Mixed	896	35.54	26.26		
Other	494	34.61	23.14		

Candidates pursued licensure at four endorsement levels: early childhood (Grades P-4), elementary school (Grades 3-8), middle school (Grades 5-9), and high school (Grades 9-12). Mean percentage learning gain scores by endorsement level (see Table 5) showed scores increasing progressively from younger-level endorsements to older-level endorsements. All comparisons were statistically significantly different ($p<.001$). Effect sizes ranged from .16 between middle school and high school to .63 between early childhood and high school.

Candidates were enrolled in either a 4-year undergraduate teacher licensure program or a 10-month MAT program. No statistically significant differences appeared in the comparison of mean percentage learning gains for the two groups.

The socioeconomic levels of schools were identified through the median gross income of residents in a school's ZIP code. All schools were then separated into quartile groups based on that statistic. Mean percentage learning gain by school socioeconomic level showed statistically significant comparisons between low-SES middle schools and all others (see Table 6). In addition, low-SES schools had statistically significantly lower mean percentage learning gain scores than high-SES

Table 4
Mean Percentage Learning Gain for P-12 Students by Learning Need

				p (Cohen's *d*)	
	n	*M*	*SD*	ELL	SPED
ELL	1,507	33.29	25.51		
TAG	1,332	37.16	27.63	.001 (.15)	.004 (.15)
SPED	1,076	33.35	24.89		
None	13,603	37.77	25.09	<.001 (.18)	<.001 (.18)

Note. ELL=English language learner. SPED=special education. TAG=talented and gifted.

Table 5
Mean Percentage Learning Gain for P-12 Students by Endorsement Level

				p (Cohen's *d*)		
	n	*M*	*SD*	Early Childhood	Elementary	Middle
Early childhood	4,061	29.40	23.43			
Elementary	5,960	35.57	24.89	<.001 (.26)		
Middle	4,555	40.67	25.51	<.001 (.46)	<.001 (.20)	
High	3,241	44.60	25.09	<.001 (.63)	<.001 (.36)	<.001 (.16)

middle schools. Effect sizes ranged from .17 (low middle-high middle comparison) to .08 (low-high middle comparison).

The content area of the unit of instruction from which learning gain scores were derived was identified for each P-12 student. Eleven hundred student scores were from units of instruction in a variety of content areas that could not be coded into the majority content areas of language arts, social studies, math, or science. Data from those units were not included in the mean percentage learning gain analysis by content area (see Table 7). Comparison of scores among the four remaining content areas were all statistically significantly different ($p<.001$). Language arts unit learning gain scores were the lowest (28.69), and social studies unit learning gain scores were the highest (45.68). Effect sizes ranged from .09 in the comparison of math and science units to .70 in the comparison of language arts and social studies units of instruction.

Candidates gathered assessment data from units of instruction completed in two semesters. The fall semester placement was the initial student teaching experience and was done at the second level of licensure endorsement in which the candidates

Table 6
Mean Percentage Learning Gain for P-12 Students by School Socioeconomic Level Quartile

				p (Cohen's *d*)	
	n	*M*	*SD*	Low	Low middle
Low	3,514	36.67	26.40		
Low middle	3,978	34.51	23.73	.001 (.09)	
High middle	2,999	38.69	25.93	.011 (.08)	<.001 (.17)
High	6,952	37.79	25.48		<.001 (.13)

Table 7
Mean Percentage Learning Gain for P-12 Students by Content Area of the Unit of Instruction

				p (Cohen's *d*)		
	n	*M*	*SD*	Language arts	Social studies	Math
Language arts	4,566	28.69	23.42			
Social studies	3,469	45.68	25.07	<.001 (.70)		
Math	3,526	38.99	24.50	<.001 (.42)	<.001 (.27)	
Science	4,549	36.66	25.51	<.001 (.32)	<.001 (.36)	<.001 (.09)

were interested. The spring student teaching experience was longer and was completed at the level at which the candidate hoped to work when hired. Table 8 shows that the mean percentage learning gain scores were statistically significantly higher in the spring (second) student teaching experience ($p<.001$). The effect size of the comparison of the two experiences was .28.

Program Changes Resulting From Data Analysis

The faculty members of the School of Education are committed to using data for continuous program improvement. Analyzing comparable data over a 5-year time period provides an opportunity to track candidate progress and programmatic decisions. The data analysis process has the potential to reveal changes over time as faculty members, adjunct instructors, university supervisors, and curricula change. It is a reality, though, that analyzing data from P-12 students is only as good as the assessment procedures used to gather those data—garbage in, garbage out.

To promote higher quality data out of this process, attention was paid to developing candidates' ability to construct assessments. Assessment instructors added instruction in the development of preassessment and summative assessment instruments and designs. Candidates were shown the previous years' learning gain results; they discussed the importance of having a matched preassessment and summative assessment and that assessments needed to be written to measure the standards-based goals in their units of instruction. University supervisors participated in sessions each fall in which there was an emphasis on monitoring candidate assessment designs. As a program, the complete assessment system was examined and redesigned to focus on student learning gains as part of meeting national accreditation standards.

The data also pointed to deficits on the part of the candidates enrolled in our 10-month MAT program in the areas of special education and technology, in addition to their struggles to write good assessments. Experiences were added to the MAT curriculum to address those deficits.

Because of the specificity of the data generated from candidate classrooms, we were able to make the most accurate assessments that we had ever been able to accomplish in the areas of diversity and SES of the classrooms and schools in which our candidates were placed. Consequently, we refocused efforts on placing

Table 8
Mean Percentage Learning Gain for P-12 Students by Unit of Instruction Semester

	n	*M*	*SD*
Fall	8,786	33.81	24.75
Spring	9,054	40.26	25.45

Note. $p<.001$; Cohen's $d=.28$.

candidates in more ethnically and economically diverse schools. Candidates now list 40% of their P-12 students in non-White (including Hispanic) categories, while the diversity of the Portland area shows approximately 27% non-White (including Hispanic) individuals. Likewise, we monitor the levels of SES of our placement schools each year.

Conclusions

The purpose of this study was to examine the impact on program improvement of systematically gathering P-12 student learning data over a 5-year period. To these ends, the most gratifying finding is that candidates can demonstrate a positive impact on student learning that is generally equivalent for P-12 students of all ethnicities and learning needs. Specifically, we identified either no statistically significant learning gain differences among P-12 students or any differences identified showed small effect sizes. These small effect sizes do not warrant major changes in program design. These data were congruent with data from our observational instruments that indicated our candidates could differentiate instruction and meet the needs of all learners.

Some findings suggest deeper investigation and will be a natural extension of this initial work. The differences in percentage learning gain scores are pronounced when compared for each of the four major content areas, showing language arts percentage learning gains that are significantly lower. This appears to be an effect of significantly higher preassessment scores for early childhood and elementary language arts students over those in other content areas. Additionally, mean percentage learning gain scores increased steadily as we examined endorsement levels with early childhood as the lowest and high school as the highest, suggesting a needed examination of assessment instruments and instructional practices. Some of these differences may be due to the forms of assessment used at each grade level. It is more typical for math and science candidates to construct assessments of 50 to 100 items, whereas early childhood candidates may conduct a performance assessment of 10 items using a 4-point rubric. For SES-level investigations, it needs to be explored why learning gains were higher among students of the middle-high socioeconomic level.

From a program point of view, the implementation and use of these assessments has had numerous positive impacts. Not only have they helped candidates learn to differentiate instruction in their classrooms but also they have provided them with substantive data to demonstrate their success in the classroom. It has been helpful that our program has data to demonstrate concrete P-12 student learning gains when our candidates are teaching as our placement director attempts to secure student teaching placements in a highly competitive market of several teacher education programs in the same geographical area. The data from the assessments have been an important part of program redesign and a focus for discussion within the fac-

ulty of program impact. The description of the process of gathering the data and examining the results assisted in supporting our assessment plan for accreditation and contributed to us to receiving Commendations in the Assessment and Diversity Standards.

As we move toward CAEP accreditation, it is important to focus even more intensely on measures of student learning. As the CAEP (2014) commission articulates, "the concept of teacher impact on P-12 student learning measures as a basis for judging preparation occurs throughout the standards, and includes measures at both the pre-service and in-service levels" (p. 22). The work of the last 5 years has produced a stable foundation for us to continue to improve our program to support our candidates and ultimately the P-12 learning that takes place in our graduates' future classrooms.

Oregon has just become an edTPA state. Thus our School of Education will be redesigning curricula and assessments to prepare candidates to pass the edTPA. One of the considerations is what will be the role of our current process of collecting data on P-12 learning gains. Initial indications are that faculty members are committed to continuing this process.

Methodologically, we realize that these measures only compose one data source in the array of multiple assessment tools that we use to understand candidate competency and program impact. But the data gathered refute voices that suggest candidate impact cannot be demonstrated in teacher preparation programs. Another concern is that placing student teachers in classrooms results in lower achievement for many P-12 subgroups. Again, these data point to other interpretations of what is happening in candidate classrooms. Candidates are able to show teachers and principals the learning gains in assessment scores that occur while the candidates are responsible for the instruction.

From a program development point of view, gathering these data has required iterative examination of the processes involved in assessing candidates and a focus on improving the quality of the assessments the candidates design and use. That work will not stop and promises to improve the quality of the data, which will allow us to make more fine-grained data evaluations. Specifically, areas we hope to improve include methods for identifying SES quartiles of schools, procedures for ensuring that candidates use matched pre- and post- measures, and procedures to accurately identify the ethnicities of the P-12 students.

There is no flawless methodology, but analyzing candidate instructional unit pretest and posttest scores provides a robust picture of candidate classrooms, and that picture has been strengthened by the force of a large data set behaving consistently over the 5 years in which we have been gathering data.

References

American Association for Colleges of Teacher Education. (2013). *Using the EdTPA*. Retrieved

from http://www.highered.nysed.gov/edtpausing.pdf

Castaiieda, M. B., Levin, J. R., & Dunham, R. B. (1993). Using planned comparisons in management research: A case for the Bonferroni procedure. *Journal of Management, 19*(3), 707-724. doi:10.1177/014920639301900311

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Psychology Press.

Council for the Accreditation of Educator Preparation. (2013). *CAEP accreditation standards*. Retrieved from http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf

Council for the Accreditation of Educator Preparation. (2014). *CAEP evidence guide*. Retrieved from http://caepnet.files.wordpress.com/2014/02/caep_evidence_guide1.pdf

Hamel, F. L., & Merz, C. (2005). Reframing accountability: A preservice program wrestles with mandated reform. *Journal of Teacher Education, 56*(2), 157-167. doi:10.1177/0022487105274458

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review, 24*(3), 328-338.

Knight, S. L., Edmondson, J., Lloyd, G. M., Arbaugh, F., Nolan, J., Whitney, A. E., & McDonald, S. P. (2012). Examining the complexity of assessment and accountability in teacher education. *Journal of Teacher Education, 63*(5), 301-303. doi:10.1177/0022487112460200

National Council for Accreditation of Teacher Education. (2013a). *Guidelines on assessment*. Retrieved from http://www.ncate.org/Accreditation/ProgramReview/GuidelinesAndProcedures/GuidelinesonAssessment/tabid/446/Default.aspx

National Council for Accreditation of Teacher Education. (2013b). *Unit standards*. Retrieved from http://www.ncate.org/Standards/NCATEUnitStandards/UnitStandardsinEffect2008/tabid/476/Default.aspx

Oregon Teacher Standards and Practices Commission. (2013). *Evidence of effectiveness* (OAR No. 584-017-1030). Retrieved from http://arcweb.sos.state.or.us/pages/rules/oars_500/oar_584/584_017.html

U.S. Census Bureau. (2014). *State and county quick facts*. Retrieved from http://quickfacts.census.gov/qfd/states/41/4159000.html