Universal Journal of Educational Research 3(10): 703-709, 2015 http://www.hrpub.org
DOI: 10.13189/ujer.2015.031007

Teacher Evaluations: Use or Misuse?

Douglas F. Warring

College of Education, University of St. Thomas, USA

Abstract This manuscript examines value added measures used in teacher evaluations. The evaluations are often based on limited observations and use student growth as measured by standardized tests. These measures typically do not use multiple measures or consider other factors in the teaching and learning process. This manuscript identifies some of the factors usually not taken into account in teacher evaluations and suggests ways to improve the teacher evaluation process. This manuscript advocates use of multiple measures in the both formative and summative evaluation of teachers and suggests the inclusion of specific student characteristics and a collegial process that needs to make the teacher evaluation process fair and more meaningful.

Keywords Teacher Evaluations, Value Added Models, Value Added Measures, Student Characteristics

1. Introduction to Teacher Evaluations

In some schools and states the concept and the use of teacher appraisal sparks discussion wherever and whenever it is mentioned. There are several questions often asked about the evaluation process. Who establishes the evaluation criteria? How is the criteria utilized? What are the desired outcomes of the evaluation process? What should the results of teacher appraisals be used for?

Education stakeholders are beginning to find some agreement in the idea that teacher appraisal can be a key factor for increasing the focus on teaching quality and continuous professional development for teachers. This belief is in keeping with the growing recognition that the quality of teaching can impact student learning outcomes. In recent years schools have been required to submit standardized assessment data of students based on the assumption that assessment data can provide credible information on progress of students, and the assessment data is often used to determine the quality of the teachers without taking other important factors into account. Investigating the connection between a teacher and his or her teaching quality has long proved methodologically challenging, largely because of the difficulty in comparing cohorts of students taught by teachers of varied experience levels with different training and backgrounds. Most people understand that one-time assessments are not a fair way to assess learning or to evaluate teachers. Well-designed evaluations systems that take into account many significant variables offer tools for growth and effective teaching. Teachers are the most influential school-based factor on student achievement [1,2,3]. Student learning growth should be measured with sophisticated statistical models and student learning is a desired outcome. Testing systems should be organized to support teachers in their efforts to improved instruction. An over reliance on standardized tests over-emphasize testing and often do not take into account the important work of teachers in teaching and student learning. Although studies have shown that some teachers are more effective than others at helping their students achieve at high levels, most indicators of teacher quality (e.g., credentials, characteristics, and observable practices) are generally poor predictors of student learning growth [4,5,6].

Teachers' scores on observation instruments have not been highly correlated with student learning growth [7]. It is not surprising that correlations are weak when the factors to be measured with observations are not well specified or when raters are poorly trained or inadequately monitored for scoring consistency after training. While overall teaching practice may be the best predictor of student achievement, classroom management is also a very significant factor.

P-12 education should also be concerned with many issues paramount to social justice, which is generally equated with the notion of equality or equal opportunity in for students and teachers in schools. Although equality is part of social justice, the meaning of social justice is actually much broader and includes opportunity and personal responsibility. The most recent theories of, and scholarly statements about social justice illustrate the complex nature of the concept, which often impacts by teacher evaluations. While student achievement is a key value or goal of schools and good teaching should be clearly identified and instruments to identify good teaching are often lacking validity. The effects of the teacher evaluation systems must be evaluated in relation to its intended impacts on teaching and learning. While value-added models (VAMs) which are the statistical

tools used to measure teacher effects on student achievement scores continue to emerge throughout districts and states across the country, education scholars simultaneously recommend caution on their use. This caution is made in reference to the inferences that are made and the evaluations and outcomes that are used based on VAM outcomes.

Student growth is often defined as the average gain in student test scores from one year to the next. It compares the test performance of a group of students in one year with the test performance of the same group of students the year before. If all students are promoted normally, student growth measures compare the test performance of a group of students in one grade with the test performance of the same group of students in the previous grade. The reality is that not all students are promoted normally, so equivalency by group by grade is fallacious and does not take into account social or cultural values or outcomes.

Systematic errors are present in attainment measures because schools serving low-achieving students are often destined to fail. This is because factors outside of the school's control that affect student learning are not taken into account in most of these measures. In growth measures, random errors are also present because when a student takes the first exam as the baseline for future progress, no one can be sure that the student being tested is putting forth maximum effort. Therefore, if all future growth is based on an inaccurate first test, then how can this measure be an accurate picture of real growth? Value-added assessment, a statistical process for looking at test score data that utilized additional factors, is one technique that researchers [5,6] have been developing to identify effective and ineffective teachers and schools.

2. Literature on Use of Standardized Tests in Teacher Evaluations

Standardized tests theoretically provide a consistent, objective means of evaluating a broad range of students on the same set of academic standards, measured in the same way. But, in a best-case scenario, they are just one piece of the puzzle of student achievement and teacher effectiveness. One of the most important questions about teacher effectiveness tied to value-added assessment is whether the estimate obtained from a value-added model can actually be called a teacher effect. Some key questions need to be addressed about teaching practices and evaluations. What changes in teaching practices are reported by teachers and documented by observational measures and student ratings? What changes occur on high-stakes achievement tests compared to the baseline year, and are these effects confirmed by independent audit tests? What is the overall impact of these on social justice? The general theory of action for test-based teacher evaluation systems holds that using student growth to measure teacher effectiveness will improve the quality of education provided to students and hence will improve student achievement. VAM is often thought of as a measure of how much a student has learned from one point in time to the next, and is often used to describe a teacher's effectiveness on a group of students' academic growth from year to year. VAM's purportedly measure a student's academic performance as a basis for determining his or her academic growth and is not related to a student's socioeconomic status or other personal characteristics that typically confound achievement-based measures.

Several school districts in the United States including Dallas, Houston, Cincinnati, Denver, New York, and Washington, D.C., and several states such as Ohio, Tennessee, and Minnesota have begun using student achievement gains as indicated by annual test scores (adjusted for prior achievement and other student characteristics) as a direct measure of individual teacher performance. These student-test-based measures are often referred to as VAM. However, even supporters of policies that make use of VAM recognize the limitations of those measures. Among the limitations are, first, that these performance measures can only be generated in the handful of grades and subjects in which there is mandated annual testing. Roughly one-quarter of K–12 teachers typically teach in grades and subjects where obtaining such measures is currently possible. Second, test-based measures by themselves offer little guidance for redesigning teacher training or targeting professional development; they allow one to identify particularly effective teachers, but not to determine the specific practices responsible for their success. Third, there is the danger that a reliance on test-based measures will lead teachers to focus narrowly on test-taking skills at the cost of more valuable academic content, especially if administrators do not provide them with clear and proven ways to improve their practice. The only definition of teacher effectiveness that seems to matter in the discussion is not comprehensive, as "increasingly, policy conversations frame teacher effectiveness as a teacher's ability to produce higher than expected gains in students' standardized test scores" [8]. Because student growth scores are relative, the evaluation system needs to guard against normative ratings that create unnecessary competition and can lead to a lack of willingness to share information with other teachers. Another significant factor is that teachers have filed suit in a half-dozen states to block complicated new evaluation formulas that in some cases have rated them based on the test scores of students they never taught. Parents have also protested that their children have been required to take tests created for the sole purpose of evaluating teachers.

According to Emma [9] Washington, DC, which was one of the first districts to incorporate student test scores in teacher evaluations, is not using those scores to rate teachers in 2015. It is pausing to give everyone a chance to get used to new exams linked to the Common Core academic standards. Maryland, New Jersey and Texas are all taking extra time to incorporate student test scores and Washington state legislators have refused to accept the administration's vision of an acceptable evaluation system, while New Mexico is adjusting its system after flawed evaluations, based on

erroneous data, caused an uproar in districts statewide.

Value-added models are statistical measures that purport to track the amount of value that a teacher adds to student learning from year to year. These are typically based on student achievement scores from different types of tests administered to students, which vary from school to school and state to state. Value-added information theoretically allows educators to assess their impact on student learning, and it can be helpful in initiating conversations about the efficacy of curriculum and instructional practices and programs. Value-added information also allows educators to better identify what is working well and potential areas for improvement to help individual students and groups of students. Above and beyond the estimates for summative evaluation, there is a wealth of diagnostic information being provided that can be appropriate for educators. National or state-level frameworks for teacher appraisal may be difficult to implement in education systems with a strong tradition of local autonomy. Education authorities need to consider different options to establish the right balance between central guidance and local flexibility. For example, if a school or local authority has already made substantial investments in building capacity for a particular teacher-appraisal framework and method, requiring it to adopt a central appraisal system may be counterproductive [10].

According to Reference [11] proponents of test-based teacher evaluations argue that growth in student achievement is the ultimate criteria for judging teacher effectiveness. They believe that value-added modeling of test-score data can do a better job of identifying the best and worst teachers compared to current indicators and that these methods are sufficiently robust in accounting for initial student differences to provide actionable data [12]. People [13,14,15] who oppose the use of VAM claim that neither standardized tests nor VAM's statistical methodology have sufficient validity for the high-stakes purpose of individual teacher evaluation and teacher pay [13].

Using value-added measures and models in teacher evaluations has become a significant issue. When it comes to improving and assessing teacher effectiveness, conditions in the school are often not taken into account. According to Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, and Shepard [14] some of these conditions include class sizes, school cultures, student mobility, area economic issues, language issues, supportive school cultures, access to needed materials, and professional development opportunities for teachers, educational assistants, and administrators. In the United States of America forty-one states require or recommend that teachers be evaluated using more than one measure, up from fifteen in 2009. Thirty-eight states require evaluations based on student achievement (eight more recommend this), and twenty-three states require indicators such as standardized tests be used for at least fifty percent of the evaluation. Across the country states have adopted new evaluation policies over the past five years, but they vary widely. However, there is more to being an effective teacher than utilizing and attempting to raise standardized test scores, yet test scores have gained widespread acceptance among the public as the key indicator of teacher performance. Most states measure a teacher's impact based on a student's academic growth or on progress compared to other students.

According to Hull [16] value-added model based on value added measures attempt to isolate the impact a teacher has on students' achievement from other factors of interest, such as student characteristics. It is important to consider the human side of teaching and learning as well as standardized test scores in measuring teacher effectiveness because teaching consists of classroom interactions among teachers and students, and teachers facilitate students' achievement of learning goals.

Evaluation systems that aligned their student learning goals with an overarching district or school goal found the goal-setting process to be clear and straightforward, offering rich and timely feedback for professional development [9]. The summative evaluations found that specialist teachers and non-teaching staff (e.g., nurses, counselors) struggled much more than did classroom teachers to adapt the student learning goals to their situations. The initial findings of this work group were that teachers view the evaluation process and this model as an effective form of professional development. Teachers and those who were involved as summative evaluators of teachers expressed hopes for increased collaboration through peer reviewer relationships. The sustainability of these models was a repeated concern of teachers and the summative evaluators, especially with the factors of coast and time. Recommendations from teachers and summative evaluators are to provide samples of completed individual growth and development plan forms, student learning goals forms, and points of contact documentation; and to clarify the relationship between goals on the individual growth and development plan and student learning goals. Critics of standardized testing often cite the loss of classroom instruction time during testing windows, which can last several weeks. Not only does this affect teachers and students being tested, but it also impacts access to the school library or computer labs, where tests are often administered As an example of these concerns an example from one of the states undergoing implementation of standards based on student test scores follows.

3. Discussion of Evaluations and Student Characteristics

A central question that must be addressed is how do we fairly account for the effect of the many types of diversity in classrooms? There is an extensive amount of attributes beyond cultural differences that must be taken into account. In a single environment, learners and teachers themselves vary in beliefs, attitudes, perceptions, self-efficacy, motivation, learning styles, cultural influences, and demographics or social identities (e.g., sex, sexual

orientation, ethnicity, ability/disability, socio-economic status, religion/spirituality, etc.). When teacher evaluations are conducted the many levels of diversity just noted (e.g., attitudes, motivation, self-efficacy, etc.) are typically not considered in calculations since data are not collected based on these factors. According to Reference [17] improvements are needed in how classroom observations are measured if they are to carry the weight they are assigned in teacher evaluation. The report's authors make specific, evidence-based recommendations aimed at improving the fairness and accuracy of teacher evaluation systems.

Under current teacher evaluation systems that do not take into account student factors, it is extremely difficult for a teacher who does not have high achieving students to receive a top rating. Teachers who have students with higher incoming achievement levels tend to receive classroom observation scores that are higher on average than those received by teachers whose incoming students are at lower achievement levels. Due to time and cost factors most schools and districts do not have processes in place to address this bias. Adjusting teacher observation scores based on student demographics is a straightforward fix to this problem. This type of an adjustment for the makeup of the class is already slightly factored into some teachers' value-added scores and should be factored into classroom observation scores as well.

The reliability of both value-added measures and demographic-adjusted teacher evaluation scores, which take into account student variability, is dependent on sample size, such that these measures will be less reliable and valid when calculated in small districts than in large districts. Thus, states should provide prediction weights based on statewide data for individual districts to use when calculating teacher evaluation scores. Observations conducted by outside observers are more valid than observations conducted by school administrators. A trained observer from outside the teacher's school who does not have substantial prior knowledge of the teacher being observed should conduct at least one observation of a teacher each year.

Using average test scores from a single year to judge school quality is unacceptable from a social justice and equity perspective. The consistent use of demographic adjustments is an unsatisfying alternative for at least two reasons. In addition to providing less accurate information about the causal impact of schools on their students' learning, the demographic adjustments implicitly set lower expectations for some groups of students than for others. This may also generate a self-fulfilling prophecy. Value-added models cannot fully control for variables because neither teachers nor their students are randomly assigned to either schools or classes, making it difficult to separate a teacher's impact on students from other non-observable measures, such as a student's motivation or help at home. The most significant finding from a Rand Corporation investigation into value-added models is that because such models might not control for all variables of interest, student achievement can never be shown conclusively to be due to individual teacher effectiveness.

Betebenner [18] found that tests were not good predictors of teacher success due to numerous issues involving student characteristics. Racism and poverty are typically not being taken into account in the teacher assessment processes currently being utilized. By privileging one way of being literate and one way of making sense of texts, the standards fail to recognize and value those students who embody various "funds of knowledge" reflecting diverse families and neighborhoods. Improvements are needed in how classroom observations are measured if they are to carry the weight they are assigned in teacher evaluation [17].

4. Fair and Equitable Evaluations

Teachers make a difference and there is a link between teacher effectiveness and student learning. VAM can be useful and the whole point of VAM is to create a more level playing field in order to make more fair comparisons among teachers. These assumptions about VAM transcend tested policy and research-based pieces. The use of VAM and its component parts have never been fully investigated or fully explained [16,17]. These assumptions, which are not fully utilized, are often ignored in order to promote VAM adoption and use by states and school districts. Policymakers and educators understand that raw achievement test scores tend to rank schools by the socio-economic status of the students served and are not fair, or consistent measures of teacher success. The very name, value-added, reflects the desire to isolate the unique contribution of schools or teachers to achievement outcomes.

The use of any value-added measure should take into account characteristics of the students and the context that affect student achievement gains. Such factors include parent education, classroom composition of special needs of students and others (e.g. English learners and special education status, poverty, homelessness), and should include consideration for student attendance, in addition to the individual student's prior achievement. This information should be taken into account both in the models and in the overall analysis of information for the ultimate evaluation judgment. Other factors that may make a significant difference include class size, the quality and availability of curriculum materials, whether students also receive tutoring or related instruction from other teachers. If these factors are not accounted for in the value-added model, they should be accounted for in the overall evaluation of a teacher. Some other approaches, with less reliance on test scores, have been found to improve teachers' practice while identifying differences in teachers' effectiveness. They use systematic observation protocols with well-developed, research-based criteria to examine teaching, including observations or videotapes of classroom practice, teacher interviews, and artifacts such as lesson plans, assignments, and samples of student work.

Discussion

Evaluation systems are difficult to develop and utilize effectively, however when they take into account student demographics they give more meaning to the career and compensation ladder for teachers by helping them to engage proactively in valuable professional development opportunities. In order to provide important feedback, value-added measures should be used only when there is a sufficient sample size and multiple years of data that take into account significant factors noted in this research report [17]. While this process is time consuming and costly, these evaluations will be more accurate than current systems of evaluation. Reference [18] found that many teachers have few students linked to them for whom data is available for both prior-year and current-year achievement. Other students who are mobile may have spent only a short time in a given teacher's classroom. Both of these are sources of considerable error. Year-to-year instability in teacher rankings is also very high. Many experts suggest that there should be at least fifty students (who have been with the teacher for a large majority of the year in each case) and at least three years of data to use in estimating a value-added score. Even with these considerations, it is important to recognize that multiple years of data may mask the year-to-year instability of scores, but do not eliminate the causes of such instability, which may often include the composition of classes that teachers teach.

The whole point of VAM is to create a more level playing field in order to make more fair comparisons among teachers. Policymakers and educators understand that that raw achievement test scores tend to rank schools by the socio-economic status of the students served and are not fair, or consistent measures of teacher success. The very name, value-added, reflects the desire to isolate the unique contribution of schools or teachers to achievement outcomes. It would be more accurate to measure student growth over a specified period of time and make allowances for highly mobile students and those who may not have started school at the beginning of the school year.

Researchers [5,17,19] found the use of any value-added measure should take into account characteristics of the students and the context that affect student achievement gains. Such factors include parent education, special needs of students (e.g., English Language Learners, special education status, poverty, homelessness), student attendance, and classroom composition, in addition to the individual student's prior achievement. In particular, studies show [6,8,16] that classroom composition greatly affects teachers' value-added scores. This information should be taken into account both in the models and in the overall analysis of information for the ultimate evaluation judgment. Other factors that may make a significant difference include class size, the quality and availability of curriculum materials, whether students also receive tutoring or related instruction from another teacher, etc. If these factors are not accounted for in the value-added model, they should be accounted for in the overall evaluation of a teacher. Among the concerns often raised by researchers [4,5,6,14] are the prospects that value-added methods can misidentify both successful and unsuccessful teachers and, because of their instability and failure to disentangle other influences on learning, can create confusion about the relative sources of influence on student achievement.

5. Conclusions

An over-reliance on standardized tests puts too much emphasis on testing and not enough on the important work of teaching and learning. Value-added measures should be used only when there is a sufficient sample size and multiple years of data. Studies find that many teachers have few students linked to them for whom data is available for both prior-year and current-year achievement. Other students who are mobile may have spent only a short time in a given teacher's classroom. Both of these are sources of considerable error. Year-to-year instability in teacher rankings is also very high because most student assessment is conducted on less than fifty students. These evaluations include students who may not have been in a class with the teacher for very long and typically including less than three years of data is used in estimating a value-added score. It is also important to recognize that multiple years of data may mask the year-to-year instability of scores, but do not eliminate the causes of such instability. The causes of instability often result from students how are in living in poverty, are homeless, or highly mobile, and often include the composition of classes that teachers teach. Districts and schools need more flexibility in developing ways to measure student performance in subjects and grades not covered by standardized tests.

The validity of teacher effectiveness ratings in any given state or district or school will depend on several factors such as the particular achievement measures used to assess the outcomes of learning, the adequacy of prior achievement data, the assignment of students to classrooms, the concurrent effects of other learning resources, the particular value added measures specifications, the quality of observational and other measures of effectiveness used in the system, and on the judgments involved in weighing evidence from multiple measures. The validity of VAMs of teacher effectiveness depends on the ability of the measures to identify and isolate teachers' contributions to their students' achievement. Existing VAMs that are currently in use differ in key aspects of their empirical specifications. This leaves policymakers with little clear guidance on what factors are important to include when constructing a fair model. At best, existing research offers insights about the potential threats to validity that need to be addressed in order to create systems for analysis and evaluation that are more fair and take into account social and cultural variables for social justice.

The end-of-year test scores do not show how much students learned that year in that class with that teacher.

Measures that take into account where students started are an improvement however, such measures of growth are only a starting point. Making judgments about individual teachers requires sophisticated analyses to sort out how much growth may be caused by the teacher and how much is caused by other factors. For example, students who are frequently absent tend to have lower scores regardless of the quality of their teacher, so it is vital to take into account how many school days students are present. Thus, to be fair and to provide trustworthy estimates of teacher effectiveness, value-added measures require complicated formulas that take into account as many influences on student achievement as possible. As previously noted, improvements are needed in how classroom observations are undertaken, measured, and used if they are to carry weight they are assigned in teacher evaluation. Researchers [5,11,17,18] make specific, evidence-based recommendations aimed at improving the fairness and accuracy of teacher evaluation systems. Key findings [5,11,17,18,19] and resulting recommendations include the fact that under current teacher evaluation systems, it is hard for a teacher who doesn't have what are considered top students to get a more favorable rating. Another finding is the fact that teachers with students with higher incoming achievement levels receive classroom observation scores that are higher on average than those received by teachers whose incoming students are at lower achievement levels, and districts do not have processes in place to address this bias. Adjusting teacher observation scores based on student demographics is a significant factor in the attempt to address this problem. Such an adjustment for the makeup of the class is already factored into some of the teachers' value-added scores and should be factored into classroom observation scores as well.

The reliability of both value-added measures and demographic-adjusted teacher evaluation scores is dependent on sample size, such that these measures will be less reliable and valid when calculated in small districts than in large districts. Thus, states should provide prediction weights based on statewide data for individual districts to use when calculating teacher evaluation scores. Observations conducted by outside observers are more valid than observations conducted by school administrators. A trained observer from outside the teacher's school who does not have substantial prior knowledge of the teacher being observed should conduct at least one observation of a teacher each year.

The inclusion of a school value-added component in teachers' evaluation scores negatively impacts good teachers in bad schools and positively impacts bad teachers in good schools. This measure should be eliminated, reduced, or revised for a more positive use in teacher evaluation systems. Collaboration with colleagues to exchange best practices and spread effective innovations and to utilize all resources (e.g., parents/families, administration, the community, and school staff other than teachers are significant factors that should be taken into account in the evaluation process. Effective evaluation systems should be based on professional teaching standards, include multi-faceted evidence, include the use of student demographic factors, use knowledgeable evaluators and a team, use evaluations that contain useful feedback connected to professional development value and encourage teacher collaboration, use expert teachers as part of the assistance and review process for new teachers and those needing extra assistance, include a panel of teachers and administrators who oversee the evaluation process, and be continually evaluated and re-designed to meet current needs and demographic changes in the student population.

Evaluation systems have an important role to play in assisting teachers to be more effective. Well-designed assessments that are formative as well as summative, are aligned with curricula, take into account student and cultural variables, are focused on higher-order skills, and with timely turnaround of results can be useful tools to support effective teaching in every subject and grade. If evaluation systems go beyond the carrot-and stick diagnostics of "good" and "bad" teachers, and instead are used as systems to support professional development, teachers and unions will be much more willing to support the evaluation reforms.

REFERENCES

[1] Rivkin, S. G., Hanushek, E. A., & Kain, J. F. Teachers, schools, and academic achievement. Econometrica, Vol.73 No.2, 417–458, 2009.

[2] Sanders, W. L., & Horn, S. P. Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for educational evaluation and research. Journal of Personnel Evaluation in Education, Vol.12 No.3, 247–256, 1998.

[3] Sanders, W. L., & Rivers, J. C. Cumulative and residual effects of teachers on future student academic achievement (No. R11-0435-02-001-97). Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center, 1996.

[4] Goe, L. The link between teacher quality and student outcomes: A research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality, 2007.

[5] Rice, J. K. Teacher quality: Understanding the effectiveness of teacher attributes. Washington, DC: Economic Policy Institute, 2003.

[6] Wayne, A. J., & Youngs, P. Teacher characteristics and student achievement gains: A review. Review of Educational Research, Vol.73 No.1, 89–122, 2003.

[7] Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. Brooklyn, NY: The New Teacher Project, 2009.

[8] Goe, L., Bell. C, and Little, O. Approaches to evaluating teacher effectiveness: A research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality, 2008.

[9] Emma, K. Rating teachers not as easy as 1, 2, 3. Politico.

Online available at http://www.politico.com/story/2014/09/rating-teachers-110467.html#disqus_thread

[10] Mead, S., A., Rotherham, & Brown, R. The Hangover: Thinking about the unintended consequences of the nation's teacher evaluation binge, Teacher Quality 2.0, American Enterprise Institute, Special Report 2, 2012.

[11] Doran, H., & Fleishman, S. Research matters/challenges of value-added assessment. Assessment to Promote Learning Vol.63 No.3, 85-87, 2005.

[12] Braun, H., Chudowsky, N., & Koenig, J. (Eds.). Getting value out of value-added: Report of a workshop. Washington, DC: The National Academies Press, 2010.

[13] Institute for Competitive Workforce. In focus: A look into teacher effectiveness. Washington, DC: US Chamber of Commerce, 2010.

[14] Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. Problems with the use of student test scores to evaluate teachers, EPI Briefing Paper #278. Washington, DC: Economic Policy Institute, 2010.

[15] Baker, E. L., & Linn, R. L. Validity issues for accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), Redesigning accountability systems for education (pp. 47-72). New York: Teachers College Press, 2004.

[16] Hull, J. Building a better education system: At a glance. Alexandria, VA: Center for Public Education, 2011.

[17] Whitehurst, G., Chingos, M., & Lindquist, K. Evaluating teachers with classroom observations: Lesson learned in four districts. Washington, DC: Brookings Institute, 2014

[18] Betebenner, D. Norm- and criterion-referenced student growth. Educational Measurement: Issues and Practice, Vol.28 No.4, 42-51, 2009.

[19] Darling-Hammond, L. Creating a comprehensive system for evaluating and supporting effective teaching. Stanford, CA: Stanford Center for Opportunity Policy in Education, 2012.