World Journal of Educational Research
ISSN 2375-9771 (Print) ISSN 2333-5998 (Online)
Vol. 6, No. 2, 2019
www.scholink.org/ojs/index.php/wjer

Original Paper

Towards the Fairest Teacher Evaluation Model

Srinivasa Rao Idapalapati[1*]

[1] English Language Center, University of Tabuk, Saudi Arabia

[*] Srinivasa Rao Idapalapati, English Language Center, University of Tabuk, Saudi Arabia

Received: February 26, 2019 Accepted: March 25, 2019 Online Published: March 27, 2019

doi:10.22158/wjer.v6n2p168 URL: http://dx.doi.org/10.22158/wjer.v6n2p168

Abstract

This study examines the fairness and reliability of the current teacher evaluation models as they are reflected in the existent literature of evaluation methods. The required data for the study are gathered from online sources. Twenty-five scholarly works on the current evaluation methods are considered for the study after a careful examination of about hundred relevant works. A qualitative analysis of the data gathered is done in light of cognitive bias theory in order to find out the major strengths and drawbacks in the models. The strengths and the drawbacks in the existent models are codified separately in the order of their occurrence in the chosen articles and brought all of them under two criteria of strengths and three criteria of drawbacks. The final set of strengths and drawbacks were analyzed quantitatively to draw inferences that are used in making recommendations for creating fairer, reliable and efficient teacher evaluation models.

Keywords

teacher evaluations, fairness of the systems, positives and negatives, guidelines for an ideal model

1. Significance of the Study

Economic growth and development of any country needs strong education systems (Clarke, 2012). The quality of teachers is the most important than any other measured aspect of schools in determining student achievement. The economical value of a good, but not great, teacher is that she/he can add up to $10600of each student’s earnings in their lifetime (Hanushek, 2011). When the teacher is given a class of 20 students, she/he can raise their aggregate earnings by $212,000.

“Improvements in teacher quality significantly reduce the probability of having a child while being a teenager, increase the quality of the neighborhood in which the student lives (as measured by the percentage of college graduates in that ZIP code) in adulthood, and raise 401(k) retirement savings rates (Chetty, Friedman, & Rockoff, 2011)”. “Accomplished teachers are worth their weight in gold, but if the profession is to be valued appropriately, it must develop standards and a system for self-evaluation (Ingvarson, 2015)". "If students have a high-performing teacher one year, they will enjoy the advantage of that good teaching in future years. Conversely, if students have a low-performing teacher, they simply will not outgrow the negative effects of lost learning opportunities for years to come (Tucker & Stronge, 2005)".

However, do we have the fair teacher evaluation systems that can support and encourage quality teachers and quality teaching? How the value or quality of a teacher is perceived? What are the bases for a decision making about a teacher? The responses to these questions as well as discussions on the responses seem a never-ending phenomenon of argument. Weisberg et al. (2009) find that the debate remains unending for two reasons. On one side, ineffective teachers are safeguarded under the shields of protective measures like tenure tracks, and on the other side, the evaluation processes neither have a defensive mechanisms to deal with discriminatory dismissals nor a protective mechanism to support, encourage and retain effective and productive teachers struggling for teacher autonomy. Tucker and Strong (2005, p. 13) hold that "if teachers do, in fact, make a difference in student learning, and if we are to have competent and caring teachers, shouldn't we relate teacher work to student work? Shouldn't student achievement be a fundamental measure of teacher effectiveness?" Kowal and Hassel (2010) suggest that designing performance measurement process needs to be based on set objectives of institutions and the expectations of institutions from the roles of every employee. When the objectives and expectations are determined institutions need to set performance standards that can be used to measure employees' performance. Once all is set, institutions need to settle on who should carry out the process and how unbiased and fair the process can be. In other fields like marketing the performance, measures can be based on tangible outcomes. For example a sales man' performance can be measured based on comparing the targeted sales quantity and achieved sales quantity. Weingarten (2010) recommends multiple reviews by the personnel with the required training, principals and evaluations by peer observers based on specific set standards of teaching, practices of teaching and student achievement. The objectives of evaluations are to improve schools and systems, to encourage and support productive teachers and to discard the unproductive teachers.

This study aims at examining the teacher evaluation systems currently in practice; find the merits and demerits in each of them and identity the most flagrant aspects in each of them with which most of the teachers acknowledge their dissension. After analyzing the findings of the exploration, we intend to suggest the educators to make use of the results of the analysis in either reducing or completely eliminating the drawbacks in the current evaluation practices or in building a new model that can ensure fairness in the evaluation process and support teacher quality as well as quality teaching of institutes.

2. Literature Review & Rationale

Toch and Rothman (2008) find that working on and realizing the construction of fair evaluation systems is as important as it is difficult. They argue that many education experts in favor of finding a solution to the fairness issue in teacher evaluation models consider students' outcome as the primary basis for grading teachers. The evaluation models, apart from providing a system that enables the decision makers in either retaining or terminating a teacher, need to suggest and arrange proper teacher training in order to improve teaching. Commenting on the prevalent confusion on the teacher accountability in higher education, the editors of Inside Higher (ed.). (2015) states that some people find the necessity to hold teachers more accountable to the students, their tuition paying parents and to the tax payers, for the quality of education and the investments in higher education. Nelson (2015) finds that the assessment models that are currently in practice habitually and almost obsessively understate the responsibility of students for their own learning, and consequentially, overstate the responsibility of teachers. In the view of Leisz kovszky (2013), schools are usually rated based on the scores of their students' performance on the standardized state tests. On the contrary, teachers' ratings are based on brief classroom observations by a principal. Murphy, Hallinger and Heck (2013) witnessed a major shift of evaluations from traits and characteristics of teachers to goals. According to their analysis of the available evidences, the teacher evaluations are moving more towards the objectives of teachers, i.e., towards the actual learning. However, the question is whether the instrument of industrial era management that privileges organizational architecture—the bureaucracy, hierarchy, and institutionalism—to direct the laboring class, the teachers, toward greater efficiency fit into the education as well?

Darling-Hammond et al. (2012) report that the current teacher evaluation systems in practice do little to support teachers in improving themselves professionally as well as in personal decision-making. They acknowledge the growing consensus on the evidences of teachers' contributions to student learning and to the teaching practices that are to be considered as part of teacher evaluation process. "For decades, teacher evaluations were little more than a bureaucratic exercise that failed to recognize either excellence or mediocrity in teaching. As such, evaluation represented a missed opportunity for giving teachers valuable feedback that could help them improve their practice" (Hull, 2013). Commenting on the classroom practices, Danielson (2012) writes that observers need a great training in order to see what actually is an evidence of teaching and not an opinion, interpretation, or bias. He says that it's not a simple matter and it's quite challenging to record and report a fact. Find and Handelsman (2012) state that

"we all like to think that we are objective scholars who judge people based entirely on their experience and achievements, but copious research shows that every one of us brings a lifetime of experience and cultural history that shapes the review process".

Weisberg, Sexton, and Keeling (2009) identify the failure of evaluation systems "to provide accurate and credible information about individual teachers' instructional performance" to the phenomenon called the Widget Effect. The Widget Effect describes the tendency of assuming classroom effectiveness to be similar from teacher to teacher. The fallacy in this view of looking at teachers is that teachers cease to be understood as individual professionals, but rather as interchangeable parts. Many scholars find that undermining the individual strengths and weaknesses of teachers and being indifferent to individual teachers' instructional effectiveness are deeply disrespectful to teachers and unmindful of the future of students.

Briggs and Domingue (2011), in their reanalysis of the data provided by Richard Budd in found significant and meaningful associations between the value-added estimates of teachers' effectiveness and their experience and educational background. Glazerman et al. (2011) report that a new generation of teacher evaluation systems seeks to make performance measurement and feedback more rigorous and useful. These systems incorporate multiple sources of information that include the metrics such as systematic classroom observations, student and parent surveys, measures of professionalism and commitment to the school community, more differentiated principal ratings, and test score gains for students in each teacher's classrooms. Obtaining information about test score gains usually incorporates a variety of statistical controls that could help in finding the differences among teachers in the circumstances in which they teach. Owing to its nature of estimating the value that individual teachers add to the academic growth of their students the kind of measure is called teacher value-added. Murphy (2012) finds that

"the shift toward results-based accountability systems is based on the idea that objective measurement of student performance is the best way to measure the performance of teachers and schools, and that associating consequences with student performance outcomes motivates better performance".

He explains that the overarching term used to describe the statistical models designed for measuring the impacts of teachers on their students' learning is Value Added Modeling (VAM). Schmoker (2013) holds that the evaluation frameworks that are in practice currently don't provide any proper solution to the teacher quality issues. Schmoker argues that the existent frameworks are to be stopped from practice for one good reason that they are bulky and confusing apart from their agenda driven language. As it was reported in *The New York Times*, some of the frame works have about 116 subcategories that should be followed in assessing a teacher. These cumbersome tools of the frameworks further deteriorate the morale of teachers that has already been smashed to a record low. On the contrary, teacher evaluations should ensure immediate improvement of the teachers by focusing and clarifying the most essential aspects of teaching to be improved, and by providing multiple opportunities to teachers to improve themselves in those aspects. The basis of teacher evaluations them would be the performance and progress of teachers on these lines.

Harris, Ingle, and Rutledge (2014) find that although it's likely that teachers with very good value-added ratings would get very good ratings from principals their study of the likelihood in the schools of a district shows a weak correlation between the principals' ratings and the value-added scores. A deeper analysis reveals that some principals give low ratings to high value-added teachers just

because they believe that the teachers' efforts in pursuing professional development opportunities are too little. On the contrary, some teachers who get high principal ratings may have got low value added-scores, and the principals tend to defend their ratings by arguing that they look, beyond the test scores of their students, for their contributions to the broader school community and consider the challenges that these teachers face viz. young children, aging parents, etc. They finally conclude that
"if high stakes are attached more to principal evaluations, then the work of teachers is likely to shift toward visible effort and social interaction with colleagues, whereas if they are applied more to value added, then teachers will probably focus more on classroom activities".

Since the information sources on teacher evaluation systems highlight the issues raised by several scholars and teachers about their findings and experiences with regard to the teacher evaluation processes, methods and frameworks that are in practice, a study of the drawbacks in order to find a solution to fix them is justifiable. And it is desirable to contribute for the improvement and refinement of the evaluation practices and so it's suggestible to other scholars as well to provide a research base for discussions in this regard.

3. Methodology and Data Analysis

The relevant data are collected by convenience sampling method based on the frequency of appearance in online sources with the assumption that the evaluation models with greater frequency of appearance are considered more important and so were selected for analysis. The data of articles considered for analysis are selected as the contents are completely relevant to the topic and have the topics that were frequently discussed and published in online sources, and are dealt and addressed well by the authors of considerable popularity in terms of the frequency of appearance of their works online. For example, "classroom observations" are a method or an aspect of a modal of teacher evaluation practices, and because it appears more frequently among the sources online, it is considered for analysis as an aspect of teacher evaluations. Then the data relevant to "classroom observations" are collected based on the relative reputation and popularity of any of the three criteria; the authors or their affiliations or their publishers. The method of sampling is purposive and convenient as they could serve well for a qualitative analysis and interpretation. Out of about one hundred articles reviewed from Google search engine, scholar Google and Questia databases twenty-five articles are chosen for analysis. The collected data are analyzed for the criteria that the scholars identified as their basis for labeling a particular aspect of a model as a drawback or an asset. All the criteria are codified into categories that are used for discussion in order to draw inferences. The inferences are used in proposing the measures required for improving and constructing better models of teacher evaluation.

4. Theoretical Framework

The natural implication of questioning the fairness of something would be the absence of fairness. As many educators hold that the current teacher evaluation models and practices leave greater scopes for biases, the question of fairness of the personnel responsible for decision making that would affect the morale and the careers of worthy teachers comes into debate. Cognitive biases can result in different ways and forms. This paper attempts to analyze and describe the flexibilities in the models that lead to biased practices in decision-making in the most popular teacher evaluation models in light of the different forms of cognitive biases as they have been described by Lee and Lebowidge (2015) in their brief article. Cognitive biases are the factors that systematically deviate the decision makers from providing rational judgments. Decision making by the people in authoritative positions becomes erratic when it is based on subjective behavioral patterns, due to which, they couldn’t make a comprehensive information search or an accurate interpretation of the information. The people with their individual behavioral patterns tend to employ cognitive heuristics in decision making that lead to a number of cognitive biases. In many cases cognitive biases influence the information noticed by individuals and disregard the negative impacts that leave on their decision making process, which usually turnout to be uncertain. Mostly the cognitive biases are due to three reasons that are overconfidence, illusion of control and belief in the law of small numbers. Over confidence refers to the abnormality of the confidence levels of the decision maker, illusion of control refers to the state of the decision makers mind that the situations are completely under their control and the law of small numbers refers to the insufficiency of the sources of information that the decision makers rely on making decisions (Simon, Houghton, & Aquino, 2000).

Kahneman and Tversky (1996), in their study of heuristics and biases approach, suggested that judgmental heuristics influence intuitive judgments and predictions. Although, the judgmental heuristics are useful they sometimes lead to characteristic errors or biases. Lee (2005) writes about the existence and prevalence of unconscious bias in today’s society. Unconscious bias is the result of developing stereotypes by the natural human processing of categorizing like objects together and relying on the stereotypical images in decision-making. Social cognitive theory identifies stereotypes as person prototypes that influence decision-making through the implicit expectancies.

5. Findings and Discussion

New Teacher Project (2010) mentions that the current evaluation systems are suffering from five major problems; they are infrequent, unfocused, undifferentiated, unhelpful and inconsequential. Infrequent evaluation of experienced teachers would leave them underperformed, as they don’t receive any meaningful feed back on their performance. The report finds that most of the evaluation models don’t focus on the most important responsibility of teachers, i.e., student academic progress, and instead they focus and rely much on the superficial judgments like teachers’ behaviors, their bulletin board presentations, etc. The current evaluation models don’t differentiate an effective teacher from an ineffective teacher and instead they treat all the teachers to be same and interchangeable parts of a wheel, which is known as widget effect. In addition, the evaluation methods don't provide the teachers with any meaningful feedback. Finally, the results of evaluation are not usually used in making important decisions like designing professional development programs and offering rewards and promotions for teachers of high performance.

There isn't any perfect evaluation system to be prescribed for all schools, and different schools opt for different evaluation models as per their choice. In spite of the existence of a number of models, the literature review for this study finds twelve evaluation models that are in practice in most of the schools. 1) Charlotte Danielson Framework for Teachers. 2) Stronger Teacher and Leader Effectiveness Performance System. 3) Mid-Continent Research for Education and Learning (McREL) Teacher Evaluation Standards. 4) Marzano's Causal Teacher Evaluation Model. 5) The Marshall Rubrics. 6) New Haven Public Schools Teacher Evaluation System. 7) DCPS's IMPACT system. 8) Houston ISD's appraisal & development system. 9) Indiana's RISE model. 10) The Rhode Island Modeland. 11) Value-added models. Mooney (2013) notes that while Charlotte Danielson Framework for Teacher is the dominant model in one of the states in the USA, the other four among the top five in the above list come in their order of preference by most of the schools in the state. The other models numbered 6 to 12 in the above list are also in practice in many of the states in the USA.

The Danielson Framework for Teaching (2011) comprises four domains of teaching: planning and preparation, the classroom environment, instruction and professional responsibilities. These domains include 22 components and 76 elements that explain the process of teacher evaluations. Murray (2014) finds that the framework "provides for a granular view when examining the effectiveness of teachers' professional practice. This instrument focuses on the provision of useful feedback, derived from this granular examination, for the improvement of teaching practices". The basis of the framework can be traced to multiple sources including Praxis III criteria that are developed by Educational Testing Services, INTASC (1992) standards and Scriven (1994). Murray's study doesn't indicate any statistical significance between the demographic variables and the items of the survey for the impact or effectiveness of the framework in changing the teaching practices. The study doesn't find any evidence in support of Framework in having an effect on instructional strategies.

The Stronge teacher evaluation model is developed into seven components; Professional Knowledge, Instructional Planning, Instructional Delivery, Assessment of/for Learning, Learning Environment, Professionalism, and Student Progress (Stronge & Tonneson, 2012). This model explains about the way teachers can collect the evidences of their teaching and what they should do as teachers before and after the class. The model identifies good teacher qualities based on the studies of top-quartile (effective) and bottom-quartile (less effective) teachers. The top-quartile teachers are decided based on their students' scores in different tests. The design of the model draws upon the findings of the cross-case studies of the qualities of top quartile teachers in light of the variables; organization, responsibility, complexity, management, verbal feedback, relationships, focus, clarity, fairness, enthusiasm, expectations, differentiation, assessment, caring, and technology (Stronge, Ward, & Grant, 2011; Popp, Grant, & Stronge, 2011).

The McREl teacher evaluation model as per one of it's successful models developed for Nortch Carolina evaluates teachers in light of six set standards; I: Teachers Demonstrate Leadership, Standard II: Teachers Establish a Respectful Environment for a Diverse Population of Students, Standard III: Teachers Know the Content They Teach, Standard IV: Teachers facilitate learning for their students, Standard V: Teachers Reflect on Their Practice and Standard VI: Teachers Contribute to the Academic Success of Students. The cycle of teacher evaluation follows, teacher Self-assessment, Professional Development Plan, Formal Observation (with pre- and post-conference), Formal Observation (with post-conference), Formal Observation (with post-conference), Peer Observation (with post-conference), Summative Evaluation Conference and finally the Summary Rating Form (NC Department of Public Instruction, 2015).

The Marzano Teacher Evaluation Model encompasses four domains into which 60 elements are accommodated. The domains viz. classroom strategies and behaviors, Preparing and Planning, Reflecting on Teaching and Collegiality and Professionalism explain the importance of supporting and developing a teacher through different means of professional development. The model focuses much on supporting teachers by providing them with the required feed back on their teaching strategies and help them build themselves on their strengths while working on correcting their drawbacks. In spite of its priority to classroom strategies and behaviors, the model emphasizes that student achievement would the ultimate goal of instruction (Marzano, Schooling, & Toth, 2010).

The Marshall Rubrics Model focuses on the evaluation of teacher performance with its rubrics categorized under six domains; Planning and Preparation for Learning; Classroom Management; Delivery of Instruction; Monitoring, Assessment, and Follow-Up; Family and Community Outreach; Professional Responsibilities. The four level rating scale that decides teachers as highly effective (4), effective (3), improvement necessary (2) and doesn't meet standards (1) doesn't have a place for student outcomes in its evaluation process and in deciding teachers' performance. Supervisors are the ultimate authority in guiding and rating teachers (Marshall, 2014).

The New Haven Public Schools Teacher Evaluation and Development System considers three major components in its process of rating teachers; student learning outcomes, teacher instructional practices, and teacher professional values. Student learning outcomes are measured by their growth on state, district, or other assessments and attainment of academic goals that are rigorous and aligned to standards. Instructional managers based on their observations of teachers evaluate the teacher instructional practices and the teacher professional values. The ratings of instructional practices are dependent on the judgments of instructional managers' observations of teachers' performances in the domains of Planning and Preparation, Classroom Practice, and Reflection. Teachers' professional values are rated by their judgments of instructional managers' observation of teachers' behaviors that address a set of characteristics including professionalism, collegiality and high expectations for students. The model lays much importance on providing intensive teacher development programs in the areas that need improvement of teachers, which are decided during the evaluation process (New Haven Public Schools, 2010).

The IMPACT is the teacher effectiveness assessment system developed and followed by the District of Columbia Public Schools (DCPSs). The aim of the system is to recognize and reward the talented and committed teachers and ensure that they are serving all the schools in the district. The model follows five steps in assessing teacher effectiveness; Clarifying Expectations, Providing Feedback, Facilitating Collaboration, Driving Professional Development and Retaining Great People. The IMPACT system considers three components in rating teachers; Teaching and Learning Framework (TLF), Teacher-Assessed Student Achievement Data (TAS) and Commitment to the School Community (CSC). Of the three components, TLF is considered for 75% of teachers' effectiveness while TAS accounts for 15% and CSC for 10%. TLF rating is dependent on formal observations conducted by the administrators, usually by principals and assistant principals, and by impartial third party observers called master educators arranged by the administrators. The observations are done in accordance with the rubrics designed for the model (National Council on Teacher Quality, 2010).

The Houston Independent School District's (HISD's) appraisal & development system strongly believes that no single measure can demonstrate a teacher's ability and effectiveness to its complete extent. Recognizing the complexity of the teaching profession the system provides teachers with multiple ways to demonstrate their success. Appraisals are based on multiple measures in three performance criteria: Instructional Practice, Professional Expectations, and Student Performance. Teachers' instructional practices and professional expectations are measured through observations by appraisers in light of a set of rubrics. The measures of student learning are based on value-added growth, comparative growth on district-wide assessments, students' progress on district-wide or appraiser-approved assessment, students' progress on district-wide or appraiser-approved performance tasks or products and, student attainment on district-wide or appraiser-approved assessments. An overall summative assessment rating is given to each teacher at the end of the year that will be used for professional developmental action plan (HISD, 2017).

The Indiana RISE Model measures teachers' effectiveness on two basic domains: Professional Practice-Assessment of instructional practice, and student learning-Contribution to student academic progress. Measure of the professional practice is done through observations in line with the Indiana Teacher Effectiveness Rubrics (TER). The measure of student learning is carrying in three different models; Individual Growth Model (IGM), School-Wide Learning measure (SWL), and Student learning objectives. For evaluation purpose the RISE models positions all teacher in three different groups; group 1 refers to those who served half of their time teaching ISTEP+ (Indiana Statewide Testing for Educational Progress-Plus) subjects. Group 2 teachers are those who taught ISTEP+ subjects less than half of their teaching period. Group 3 teachers are those who haven't taught the ISTEP+ during their teaching period. When it comes to summative rating of teachers' effectiveness the group 1 teachers' rating are dependent on 50% of the scores on their professional practice and 50% on their student academic progress. For group 2 teachers their rating depends 60% on their professional practice and 40% on their student growth while the ratings of group 3 teachers is dependent on 75-25% of their professional practice and student growth (Indiana Department of Education, 2010).

The Rode Island Teacher Evaluation model works on four major components in its evaluation process—professional practice: Classroom Environment, Professional Practice: Instruction, Professional Responsibilities, and Student Learning. The final ratings are based on the scores of each teacher in the four components on a 400-point scale. Of the four major components student leaning is considered for 30% and the other components are considered for 70% with 25% for classroom environment, 25% for instruction and 20 percent for professional responsibilities (RIDE, 2015).

A critical analysis of the popular teacher evaluation models, in light of the available literature and the cognitive bias theory, suggests that the prominent drawbacks in the models are due to the biases of the evaluators since there is no perfect evaluation system that enables teachers to demonstrate their efficiency in the fairest terms. Although a great number of researches recommend considering student achievement as the major criterion in rating teachers, most of the evaluation models rest below 50% of their evaluation-rating stake on student achievement. Leaving the major stakes of teacher evaluation ratings on the observation reports can result in the constriction of teacher autonomy and forces teachers to think always about the observation rubrics at the cost of reflective and creative teaching practices. The cognitive bias theory suggests that irrespective of the number of years of experience and the amount of training, the personnel in the decision making positions can never go unbiased in rating teachers (Simon, Houghton, & Aquino, 2000).

A strong link between student evaluations and teacher evaluations seems to be that teachers' evaluations need to be dependent on their students' evaluations. Student evaluations are either formative or summative. Formative evaluations are in the form of quizzes, face to face interviews and daily/weekly/monthly tests, projects, assignments, etc. Summative assessment is in the form of final tests conducted by either district or state or central boards. Clarke (2012) finds a strong link between high quality formative assessment activities and better learning outcomes as they are measured by student performance on standardized tests of educational achievement. Clarke mentions three kinds of student assessment methods viz. classroom assessments, examinations, and large-scale system-based assessments. While classroom assessment are usually formative and are meant for teachers' understanding of their students' progression periodically they aren't determinants of students' achievement in a year to be used in recruitments for further studies. The results in the examinations and the system-based assessments reflect students' annual achievement or learner outcomes. However, technical quality of the tests determines whether the accountability of these test-based assessments is positive or negative. Stuggins et al. (2004) categorize all methods of student assessments into four groups: Selected response and short answer, Extended written response, Performance assessment and Personal communication.

Literature review relevant to this study further clarifies that formative assessments are low-stakes assessments that are meant for teachers' and institutions' estimation of learners' progress and for using the results in planning future course of action in teaching and training. On the other hand, summative assessments are high-stakes assessments used for awarding the final grades and scores. Brown, Bull, and Pendlebury (2013) specify that students' learning styles depend much on assessment methods. It's the assessment method that influences the students in identifying the important things to learn instead of giving importance to what the teacher says as important. They mention, "If you want to change student learning then change the method of learning". Citing Whatkins and Hatties (1985), they mention that testing student achievement/learning through multiple choice questions and other forms of tests promote reproductive style of learning and testing based on projects and open-ended questions can promote autonomous and deeper strategies of learning. Finally, the measure of a student's achievement in a year is reflected in the form of a grade and a scorecard. It's the high stakes students' evaluations that need to be considered in high stakes teacher evaluations.

Daley and Kim (2010) describe the Value-added assessment as a method of measuring the contribution of teachers and schools to the growth of their students' academic achievement during a school year. By this method, the individual growth of students is measured by matching each student's test scores to his or her previous scores. Value-added assessment helps in separating the impact of a school year on a student's learning from the student's prior experiences in and out of school, as well as the student's individual characteristics such as demographics, socioeconomic status, and family conditions. Many scholars find that value added evaluation models usually attribute the credit of the student achievement to individual teachers and do not consider the external factors that influence students' scores. They maintain that the impact of the factors like the size of the classes, instructional time, availability of specialists, tutors, availability of resources like books, computers, science labs, home and community supports, individual student needs, health, attendance, peer culture, prior teachers and schooling, other current teachers, differential summer learning loss and the specific tests used that emphasize some kinds of learning are ignored in the value added models (Darling-Hammond et al., 2012). The statement of American Statistical Association (2014) reports

"VAMs may provide quantitative information that is relevant for improving education processes. For example, the models can provide information on important sources of variability, and they can allow teachers and schools to see how their students have performed on the assessment instruments relative to students with similar prior test scores. Teachers and schools can then explore targeted new teaching techniques or professional development activities, while building on their strengths".

During the codification process initially, the teacher evaluation issues with regard to the teacher performance were given the code A the issues relevant to the principal involvement were given code B and the issues relevant to teacher observers were listed under code "C". On the other hand, the voices that consider students' scores in the teacher evaluations were listed under the code D. A qualitative analysis of the codified issues that were present in most of the evaluation models reveals that the models are in a broad sense can be divided into two categories. One category is that the models that support observations of teachers' performance by principals or peers or outsiders hired by the institutes and the other one is of the models that support and consider student achievement. In the next stage, the arguments in favor of teacher observations either by principals or observers or by both are brought under A1 and the opinions that support teacher independence and student achievements were given the code A2. In the next stage of the codification, based on the number of opinions in support of each specific issue that can the categorized under either A1 or A2 the overall criteria for strengths and drawbacks are charted.

The codification process for all the strengths boils down to the following two criteria.

1) Continuous feed back-the continuous and immediate feedback of observers helps teachers reflect on their instructional strategies.

2) Identifying drawbacks in the instructional methods- Skillful and trained observers can identify the major drawbacks in the instructional practices of teachers that can help in designing professional development programs to improve the teacher quality of teachers.

The codification process brings forth the following criteria of drawbacks.

1) Missing student achievement element—Except for the value added models that are not adopted by most of the schools; other prominent models don't consider student achievement as a major element to be considered in the teacher evaluation process.

2) Lengthy observation rubrics—the set of rubrics designed for observers is too lengthy for any teacher to focus and follow, which proves to be distractive to the teachers from being innovative and productive.

3) Relying more on the observers and their reports-Major stake of teacher quality rating is dependent on observers' report.

Table 1. The Final Criteria of Strengths and Drawbacks after the Codification Process

Criterion	No of Articles considered as a Strength (n=25)	No. of Articles Considered as a Setback (n=25)
Continuous Feed back	23	0
Identifying drawbacks in teachers' instructional methods for the purpose of providing relevant training programs.	16	0
Missing Student achievement element/lesser importance to student achievement	4	21
Lengthy observation rubrics	3	22
Relying more on the observers and their	9	16

reports

An analysis of the details in the table above leads to the inference that most of the studies are in in favor of the criteria of having continuous feedback, regular training programs, considering student achievement as a major factor in teacher assessment, limited observation rubrics and minimizing the factor of observations and their reports. The following table is the result of the inferences based on Table 1, and it demonstrates the required criteria for a fairer teacher evaluation model.

Table 2. The Criteria that Are Preferred in Fairer Teacher Evaluation Models

Criterion	No of Articles in favor (n=25)	No. of Articles against + neutral (n=25)
1. Continuous Feed back	23 (92%	0 (0%)
2. Identifying drawbacks in teachers' instructional methods for the purpose of providing relevant training programs.	16 (64%)	0 (0%)
3. Student achievement element as a factor of teacher assessment	21 (84%)	4 (16%)
4. Fewer observation rubrics	22 (88%)	3 (12%)
5. Observers and their reports not as a major factor in the teacher assessment	16 (64%)	9 (36%)

Table 3. Quantitative Analysis of the Details in Table 2

	Mean	Mode	Median	Standard Deviation
Articles in favor of the criteria in Table 2	19.6	16	21	3.36
Articles Against + Nuetral to the Criteria in Table 2	3.2	0	3	3.70

6. Conclusion

The forewords of all the evaluation models state that the primary aim of teaching and the teacher evaluation models would be to improve the learning outcomes of students. Most of the evaluation models that we discussed in this work are developed and practiced in the United States. The models have been an improvisation of their antecedent models that were considered to be traditional in the sense that they didn't look at teacher quality in terms of the learner outcomes. Most of the traditional models were criticized for their ineptitude in encouraging and improving teacher quality because of their nature of leaving the choice and scope of decision making, in the issues of teacher recruitments, termination, sanctions and retentions, basically on the institutional leaders or heads like principals, directors or the other administrative personnel with different designations. In spite of all the basic aims of the reformation initiatives, the models still leave a major role of decision making in the important issues, concerned with recognizing and supporting effective teachers, on the institutional leaders allowing the ages old crisis continue in different forms. The models find different excuses for supporting and retaining the hegemony of the bureaucratic managerial strategies that incessantly strive to constrict teacher autonomy just for the sake of satisfying the bossy attitudes of the people in the decision making positions. The study finds that the concept of observations of teachers' classroom practices and classroom behaviors is the most important tool of the bureaucratic managerial machine that would account for more than 50% of the teacher evaluation stakes that in turn leave a greater scope for administrative personnel to target, persecute and provide a biased rating for teachers from their personal perspective of teacher quality and can brand an effective teacher as an ineffective. The inferences of the present study suggest that the teacher evaluation models would be more effective in encouraging teacher quality and in retaining and supporting effective teachers if they can orient their stakes in rating teachers toward student learning outcomes that can be measured and projected in the form of scores. The evaluation models that rely more on a specific teacher's impact on their students' achievement for more than 75% of evaluation rating stake can improve teacher quality. At the same time, the models shouldn't ignore the importance of constructive feed back of observers that can be used well in designing profession developmental programs as well as in placing the teachers on the wheels of reflective teaching practices.

Recommendations

Based on the inferences drawn from the resultant criteria of strengths and drawbacks that could impact teachers' performance, their productivity and their career, and in view of contributing to the future evaluation models, it is recommended that teachers are to be trained well for being well aware of the institutional goals that would be in compliance with national educational goals. With the objective of reaching the set goals that are national and institutional with regard to their students, the institutes need to take the responsibility of supporting their teachers by providing regular in service professional development programs for upgrading and improving teacher quality. With the required teacher training programs keep continuing regularly, teachers are to be set free with regard to their work in classes as well as in dealing with their students. It is highly recommended that the classroom observations by peers, principals and external experts need to be confined to providing feedback on the performance of the teachers and they should not have any role in making decisions about teacher retentions or terminations in the final annual teacher appraisals. The decisions about the teacher retentions, promotions and terminations need to be based more on the student achievements and teachers' productivity that are usually reflected in terms of grades and scores in tests. Observer evaluations in combination with student evaluations can be considered for about 15% of their teachers' quality in making decisions about promotions, retentions and terminations. can be used to some extent when with the given more freedom of choice in dealing with students. In brief, the teacher evaluation models need to provide greater scope for teacher autonomy as they are the creative classroom performers and need to be kept away from the subjective evaluations of any others by supporting them with relevant and regular professional development programs and by providing them with regular and continuous feedback.

References

American Statistical Association. (2014). ASA statement on using value-added models for educational assessment. *Alexandria*. Retrieved from http://www.amstat.org/policy/pdfs/asa_vam_statement.pdf

Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational measurement: Issues and practice*, *11*(1), 36-44. https://doi.org/10.1111/j.1745-3992.1992.tb00230.x

Assessment, I. T., & Support Consortium. (2011). In *TASC model core teaching standards: A resource for state dialogue*. Washington, DC: Council of Chief State School Officers.

Bangert, A. (2006). *Two traditions for assessing student achievement*. Retrieved from http://www.montana.edu/teachlearn/Papers/tradassess.pdf

Briggs, D., & Domingue, B. (2011). Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the "Los Angeles Times". *National Education Policy Center*. Retrieved from http://files.eric.ed.gov/fulltext/ED516008.pdf

Brown, G. A., Bull, J., & Pendlebury, M. (2013). *Assessing student learning in higher education*. Routledge. Retrieved from https://www.coursehero.com/file/p6tl9vv/Brown-G-A-Bull-J-Pendlebury-M-2013-Assessing-student-learning-in-higher/

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (No. w17699). National Bureau of Economic Research. https://doi.org/10.3386/w17699

Clarke, M. (2012). *What matters most for student assessment systems: A framework paper*. Retrieved from https://www.openknowledge.worldbank.org/bitstream/handle/10986/17471/682350WP00PUBL0WP10READ0web04019012.txt?sequence=2

Daley, G., & Kim, L. (2010). *A Teacher Evaluation System That Works*. Working Paper. National Institute for Excellence in Teaching. Retrieved from http://files.eric.ed.gov/fulltext/ED533380.pdf

Danielson, C. (2012). Observing classroom practice. *Educational Leadership*, *70*(3), 32-37. Retrieved from http://cms1.cms4schools.com/nbexcellence/cms_files/resources/Observing%20Classroom%20Practice.pdf

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan,* 93(6), 8-15. https://doi.org/10.1177/003172171209300603

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 8-15. https://doi.org/10.1177/003172171209300603

Evans, B. R., Wills, F., & Moretti, M. (2015). Editor and Section Editor's Perspective Article: A Look at the Danielson Framework for Teacher Evaluation. *Journal of the National Association for Alternative Certification*, *10*(1), 21-26. Retrieved from http://www.jnaac.net/index.php/test/article/view/144

Fine, E., & Handelsman, J. (2012). *Reviewing Applicants: Research on Bias and Assumptions*. WISELI, University of Wisconsin-Madison. Retrieved from http://wiseli.engr.wisc.edu/docs/BiasBrochure_3rdEd.pdf

Glazerman, S., Dan, G., Susanna, L., Stephen, R., Douglas, O. S., Grover, J., Whitehurst, & Michelle, C. (2011). Passing muster: Evaluating teacher evaluation systems.

Hanova Research. (2012). *Best Practices for Including Multiple Measures in Teacher Evaluations*. Washington DC. Retrieved from http://www.hanoverresearch.com/wp-content/uploads/2012/05/Best-Practices-for-Including-Multiple-Measures-in-Teacher-Evaluations-Membership.pdf

Hanushek, E. A. (2011). Valuing Teachers: How Much is a Good Teacher Worth? *Education Next*, *11*(3), 40-45. Retrieved from http://hanushek.stanford.edu/publications/valuing-teachers-how-much-good-teacher-worth

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How Teacher Evaluation Methods Matter for Accountability A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures. *American Educational Research Journal*, *51*(1), 73-112. https://doi.org/10.3102/0002831213517130

Heather, F., Ketteridge, S., & Marshall, S. (2003). *A handbook for teaching & learning in higher education: Enhancing Academic Practice*. Routledge. New York and London.

HISD. (2017). *Teacher Appraisal and Development System: Survey Analysis, 2016-2017*. Retrieved from https://www.houstonisd.org/teacherappraisal

Hull, J. (2013). *Trends in Teacher Evaluation: At a Glance*. Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/teacherevalreview

Indiana Department of Education. (2010). *RISE Evaluation and Development System*. Retrieved from https://www.doe.in.gov/sites/default/files/evaluations/rise-handbook-2-0-final.pdf

Ingvarson, L. (2015). *Recognizing the value of teachers*. Research Developments, ACER. Retrieved from http://rd.acer.edu.au/article/recognising-the-value-of-teachers

Inside Higher (ed.). (2015). *New Debates on Accountability*. Inside Higher Ed. Washington DC. Retrieved from https://www.insidehighered.com/system/files/media/New%20Debate%20About%20Accountability.pdf

Jackson, T., Burrus, J., Bassett, K., & Roberts, R. D. (2010). Teacher leadership: An assessment framework for an emerging area of professional practice. *ETS Research Report Series*, *2010*(2), I-41. https://doi.org/10.1002/j.2333-8504.2010.tb02234.x

Kahneman, D., & Tversky, A. (1996). On the Reality of Cognitive Illusions. *Psychological Review, 103*(3), 582-591. Available on American Psychological Association, Inc. https://doi.org/10.1037/0033-295X.103.3.582

Kowal, J., & Hassel, E. A. (2010). *Measuring Teacher and Leader Performance: Cross-Sector Lessons for Excellent Evaluations*. Building an Opportunity Culture for America's Teachers. Public Impact. Chapel Hill, NC. Retrieved from http://www.publicimpact.com/images/performance_measurement-public-impact.pdf

Lee, A. J. (2005). *Unconscious bias theory in employment discrimination litigation* (No. 40, p. 481). Harv. CR-CLL Rev.

Lee, S., & Lebowitz, S. (2015). 20 cognitive biases that screw up your decisions. *Business Insider*. Retrieved from http://www.businessinsider.com/cognitive-biases-that-affect-decisions-2015-8

Leiszkovszky, I. (2013). *Grading the Teachers: Measuring Teacher Performance Through Student Growth*. State Impact. Retrieved from http://stateimpact.npr.org/ohio/2013/06/17/grading-the-teachers-measuring-teacher-performance-through-student-growth/

Levin, B. (1979). Teacher Evaluation—A Review of Research. *Educational Leadership*, *37*(3), 240-45.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing Critical Thinking in Higher Education: Current State and Directions for Next‐Generation Assessment. *ETS Research Report Series*, *2014*(1), 1-23. Retrieved from http://onlinelibrary.wiley.com/store/10.1002/ets2.12009/asset/ets212009.pdf?v=1&t=ig3m1co2&s=85564eaae960414af92fce4d2d09880407bedf0a

Marshal, K. (2014). *Teacher Evaluation Rubrics*. Retrieved from http://usny.nysed.gov/rttt/teachers-leaders/practicerubrics/Docs/marshall-teacher-rubric-jan-2014.p df

Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan*, *86*(10), 727-735. https://doi.org/10.1177/003172170508601004

Marzano, R., Schooling, P., & Toth, M. (2010). Creating an aligned system to develop great teachers within the federal Race to the Top initiative. *Marzano Research Laboratory*. Retrieved from http://www.marzanoevaluation.com/files/Marzano-Race-to-the-Top-White-Paper.pdf

Mooney, J. (2013). *Majority of NJ Schools Opt for Widely Used Teacher-Evaluation Method*. NJSpotlight, Education, February. Retrieved from http://www.njspotlight.com/stories/13/02/21/majority-of-nj-schools-opt-for-widely-used-teacher-evaluation-method/

Murphy, D. (2012). *Where is the Value in Value-Added Modeling*? Nanyang Polytechnic, SINGAPORE. Retrieved from http://images.pearsonassessments.com/images/tmrs/Where_is_the_Value_in_Value-Added_Modeling.pdf

Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via Teacher Evaluation The Case of the Missing Clothes?. *Educational Researcher*, *42*(6), 349-354. https://doi.org/10.3102/0013189X13499625

Murphy, R. (2013). Testing teachers: What works best for teacher evaluation and appraisal. *The Sutton Trust*. Improving the Social Mobility Through Education. Retrieved from http://www.suttontrust.com/wp-content/uploads/2013/03/MURPHYTEACHEREVALUATION-FINAL.pdf

Murray, P. K. (2014). *An Investigation of Teacher and Administrator Perceptions of Pennsylvania's New Teacher Evaluation System, Based Upon the Danielson Framework For Teaching, and Its Impact on Teachers' instructional Strategies in an Urban School District* (Doctoral dissertation, Indiana University of Pennsylvania).

National Council on Teacher Quality. (2010). *IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel 2010-2011*. Retrieved from https://www.nctq.org/publications/IMPACT:-The-District-of-Columbia-Public-Schools-Effectiveness-Assessment-System-for-School--Based-Personnel-2010--2011

NC Department of Public Instruction. (2015). *North Carolina Teacher Evaluation Process*. NC State Board of Education, NC.

Nelson, H. B. (2015). Assessing Assessment. *New Debates on Accountability*. Inside Higher Ed. Washington DC. Retrieved from https://www.insidehighered.com/system/files/media/New%20Debate%20About%20Accountability.pdf

New Haven Public Schools. (2010). *NHPS Evaluation and Development System Recommendations*. Retrieved from http://www.nhps.net/node/2328

Popp, P. A., Grant, L. W., & Stronge, J. H. (2011). Effective Teachers for At-Risk or Highly Mobile Students: What are the Dispositions and Behaviors of Award-Winning Teachers?. *Journal of Education for Students Placed at Risk (JESPAR)*, *16*(4), 275-291. https://doi.org/10.1080/10824669.2011.610236

RIDE. (2015). *Rode Island Model Evaluation and Support System*. Retrieved from http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Guidebooks-Forms/Teacher_Guidebook_2015-16.pdf

Schmoker, M. (2013). Why complex teacher evaluations don't work. *Colleagues, 10*(1), 5. Retrieved from http://scholarworks.gvsu.edu/colleagues/vol10/iss1/5

Scriven, M. (1994). Duties of the teacher. *Journal of Personnel Evaluation in Education*, *8*(2), 151-184. https://doi.org/10.1007/BF00972261

Simon, M., Houghton, S. M., & Aquino, K. (2000). Cognitive biases, risk perception, and venture formation: How individuals decide to start companies. *Journal of business venturing*, *15*(2), 113-134. https://doi.org/10.1016/S0883-9026(98)00003-2

Stiggins, R., Arter, J., Chappuis, J., & Chappuis, S. (2004). Assessment Methods. *Classroom Assessment for Student Learning: Doing it Right-Using it Well*. Assessment Training Institute. Retrieved from http://www-tc.pbs.org/teacherline/courses/inst325/docs/inst325_stiggins.pdf

Stronge, J. H., & Tonneson, V. C. (2012). *Stonge Teacher Effectiveness Performance Evaluation System*. Stronge and Associate Educational Consulting, LLC. Retrieved from http://www.mcvea.org/extras/StrongeBook.pdf

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339-355. https://doi.org/10.1177/0022487111404241

Sundberg, M. D. (2002). Assessing Student Learning. *Cell Biology Education*, *1*, 11-15. https://doi.org/10.1187/cbe.02-03-0007

The New Teacher Project. (2010). *Teacher Evaluations 2.0*. Retrieved from http://tntp.org/publications/view/teacher-evaluation-2.0

Toch, T., & Rothman, R. (2008). Rush to Judgment: Teacher Evaluation in Public Education. Education Sector Reports. *Education Sector*. Retrieved from http://eric.ed.gov/?id=ED502120

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. ASCD.P. 5 Retrieved from http://www.ascd.org/publications/books/104136.aspx

Wakeford, R. (2003). Principles of student assessment. In *A Handbook for Teaching & Learning in Higher Education* (pp. 42-61). London, UK: Kogan Page Ltd.

Washington: Brooking Institution. (2011). 1-36. Retrieved from https://www.researchgate.net/profile/Dan_Goldhaber/publication/254429689_Passing_Muster_Evaluating_Teacher_Evaluation_Systems._Washington_DC_Brown_Center_on_Education_Policy_at_Brookings/links/02e7e538f05d035be5000000.pdf

Watkins, D. A., & Hattie, J. (1985). A longitudinal study of the approach to learning of Australian tertiary students. *Human Learning*, *4*, 127-142.

Weingarten, R. (2010). A New Path Forward: Four Approaches to Quality Teaching and Better Schools. *American Educator*, *34*(1), 36-39.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. *The New Teacher Project*. Retrieved from http://widgeteffect.org

Wood, D. F. (2003). Problem based learning. *BMJ : British Medical Journal*, *326*(7384), 328-330. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1125189/pdf/328.pdf