*The Qualitative Report* Volume 15 Number 6 November 2010 1504-1522
http://www.nova.edu/QR/QR15-6/namaghi.pdf

# A Data-Driven Conceptualization of Teacher Evaluation

Seyyed Ali Ostovar Namaghi
Shahrood University of Technology, Shahrood, Iran

*Research perspectives on teacher evaluation present evaluators with a set of possible acts. Local evaluation systems, on the other hand, specify a permissible set of acts from the total universe. The acts specified within a given locality act as conditions for teacher action. Using the sampling and analytical procedures of grounded theory, this study aims at exploring how evaluation of teaching performance in universities of Iran conditions practitioners' action (**conditions**), what teachers do in the face of these conditions (**action**), and the effect these conditions and actions have on practitioners' professional life (**consequences**). The findings will be useful for stakeholders since they show the other side of the teacher evaluation coin: one side being the research perspectives while the other being practitioners' perspectives.*
*Key Words: Teacher Evaluation, Teachers' Perspectives, Grounded Theory, and Local Evaluation Systems*

## Introduction

Danielson and McGreal (2000) stated two primary purposes of teacher evaluation: quality assurance and professional development. The former is achieved through summative evaluation while the latter is achieved through formative evaluation. Summative evaluation aims to license, hire, give tenure to, promote, demote, or dismiss teachers. On the other hand, formative evaluation aims to encourage the professional growth and development of its teachers, shape performances by giving appropriate feedback, build new practices or alter existing practices (Peterson, 2000).

Although both types of evaluation aim to measure teacher performance, the formative evaluation identifies ways to improve performance and the summative evaluation determines whether the performance has improved sufficiently such that the teacher can be rewarded. While each type is valuable, neither type can lead to reform on its own. When coupled, however, formative and summative evaluations provide optimal professional development opportunities (see Nolan & Hoover, 2005) and tenure (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007).

Despite their complementary nature, some teacher evaluation systems focus on summative evaluation at the cost of formative evaluation. They use summative evaluation to build a case to dismiss incompetent teachers. This approach has several drawbacks: (a) it is not conducive to fostering an honest, open, and pedagogically sophisticated dialogue between principals and teachers; (b) it raises the level of tension and anxiety and makes it more difficult to admit errors, listen, and talk openly about areas that need improvement; (c) it doesn't prod teachers to emerge from their isolation and reflect with their colleagues on what they need to change in order for more students to succeed; (d) it doesn't give clear direction on the ways in which teachers can improve their performance; and finally it does not motivate a mediocre teacher to improve — or spur a good teacher on to excellence (Danielson & McGreal, 2000).

Crew, Everitt, and Nunez (1984) found two major philosophical problems with judgmental evaluation. First, it focuses on poor teacher performance and gathers documentation on a teacher's weaknesses. Second, it does not candidly address weaknesses observed in teachers. Evaluation will be more conducive to thought and reform if it focuses on the positive side of teacher action.

Evaluation will be exempt from the foregoing pitfalls if it systematically links teacher evaluation and staff development (Marshall, 2005). Marshall believes that evaluation facilitates teacher growth if it is based on multiple sources of data, includes clear, relevant, and meaningful performance criteria, focuses on peer assistance and teacher goal setting, and fosters mutual trust between the teacher and evaluator.

Evaluation can be limiting if it is judgmental. It can be limited if it is based on a single source of data. For instance, in some universities such as Iranian universities, evaluation is mainly based on students' views. To provide a better picture of teaching performance, students' views should be juxtaposed to the review of teachers' lesson plans (Stronge, 2007), classroom observations (Mujis, 2006), self-assessments (Uhlenbeck, Verloop, & Beijaard, 2002), portfolio assessments (Brandt et al., 2007), student achievement data in standardized tests (Brandt et al.), and student work-sample reviews (Mujis). Though useful as a source of information, each of the foregoing methods of teacher evaluation has its own limitations.

1. Review of lesson plans: planning is a window to teacher preparation and correlates with student learning (Stronge, 2007), but lesson plans are adjusted during their implementation. Thus, assessment of plans cannot account for the quality and appropriateness of adjustments.
2. Classroom observations: observation captures information about what actually occurs in the classroom (Mujis, 2006), but poorly trained observers and brief observations are usually biased (Shanon, 1991).
3. Self-assessment: reflection or teachers' retrospective analysis of instruction encourages teachers to learn (Uhlenbeck et al., 2002); though useful, however, it demands time and administrative support. Hence its use is contingent upon administrators' priorities (Peterson & Comeaux, 1990).
4. Assessment of portfolios: portfolios help evaluators to identify strengths and weaknesses. They also encourage professional development (Attinello, Lare, & Source, 2006). Despite their usefulness, they should be used cautiously because there are no conclusive findings on their reliability. Another concern is their practicality, i.e., the required time to develop and review portfolios (Tucker, Stronge, & Gareis, 2002).
5. Student achievement data: the use of standardized test scores enables evaluators to measure the efficiency of instruction. But the problem is that such tests are not available in some education systems. Moreover, these tests measure only a portion of the syllabus and teachers' effects on learning (Berry, 2007).
6. Student work-sample reviews: in comparison with standardized tests, student work samples may help to better identify what aspects of teaching relate to student learning (Price & Schwabacher, 1993). But the main problem is that they can be

very time-consuming. Moreover addressing the issues of validity and reliability proves difficult for the evaluator.

To summarize, available theoretical perspectives help us avoid two problems in teacher evaluation practices: (a) emphasis on summative evaluation to judge and dismiss, and (b) reliance on one single source of data. Such practices are limiting because they do not lead to professional growth. They are limited because they ignore many alternative sources of data. The literature reviewed also suggests that quality teacher evaluation aligns not only the summative and formative function of evaluation, but also presents a fuller picture of teacher performance by relying on multiple sources of data. Though promising, reforming evaluation based on theoretical perspectives and research findings is limited since it ignores a very important source of data: the perspectives of those who are evaluated. Thus quality evaluation depends on accommodating not only researchers' perspectives but also practitioners' perspectives. And it is this latter source of data that this article seeks to explore.

### Purpose of the Study

To reform any teacher evaluation system, the evaluator should be informed by two sources of data: (a) theoretical perspectives and research findings, and (b) practitioners' perspectives. With this insight and through elaborate coding schemes of grounded theory, this study aims at developing a data-driven conceptualization of teacher evaluation by exploring: (a) the socially-given or local teacher evaluation criteria (Conditions); (b) practitioners' perspectives and actions in the context of these criteria (action); and (c) the effect on their professional lives of these criteria (Consequences).

## Research Method

### Research Context

This study was conducted at Shahrood University of Technology (SUT) in Iran. In this context practitioners are evaluated by a teacher evaluation tool consisting of 15 items. The tool is general and as such it does not measure aspects specific to any given course. Despite the promising sources of teacher evaluation data, evaluators in this context make use of one source, i.e., students' evaluation of teaching performance. Having noticed the shortcomings of the present approach, evaluators intend to collect data from two additional sources in the future: colleagues' review of teaching performance, and the views of students with higher GPAs. Evaluation results are mainly used for promotion and giving tenure.

Having taught English as a foreign language (EFL) for five years in this context, the researcher had an insider's view of teacher evaluation in this university. Informal conversation with colleagues from different departments at recess presented the researcher with a deep theoretical sensitivity about teacher evaluation. This sensitivity motivated the researcher to write a proposal and submit it to the research department at SUT. This department approved and funded the project.

## Data Collection

Following Strauss and Corbin (1998), the researcher theoretically sampled concepts related to teacher evaluation from interview data. Following Seidman (1991) he designed interviews to acquaint the participants with the nature of the study, to establish rapport, to set a context for the phenomenon under study, and then to obtain deep and detailed descriptions of the experience. The study started with the general question, "How do you evaluate teacher evaluation at SUT?" Analysis of preliminary data revealed the theme of dissatisfaction. Having uncovered areas of dissatisfaction, the researcher collected more data to uncover the determining conditions of dissatisfaction. To move beyond description and explanation, the researcher tried to uncover the consequences of the current evaluation scheme. Thus data collection and analysis aimed at acquiring descriptive, explanatory and predictive power for emerging concepts and categories.

Instead of statistical sampling that starts with a representative sample of participants, the researcher focused on theoretical sampling by selecting subsequent subjects based on the information that emerged from the data already coded. Having interviewed twelve probationary and tenured practitioners with a minimum of six years of teaching experience, the researcher stopped sampling since theoretical saturation was reached. Following Brown (1999) this type of purposive sampling aimed at increasing the diversity of the sample and the richness of the concepts and categories.

## Data Analysis

Interviews were transcribed to best represent the dynamic nature of the living conversation. Each of the verbatim transcripts was returned to the participant for his review so he could remark on the accuracy of the document. During the research, each participant was assured confidentiality through the use of concepts rather than names in the reporting of data. They were also assured that once the data are coded, connection back to the individual participant is almost impossible to trace. Identification of the individual participant is not paramount, because the concepts generated by the participants—not the individual participants—are at the centre of study (Glaser, 1978).

Transcribed interviews were open coded to conceptualize and categorize data. This was achieved through two basic analytic procedures. Once categories were formed in open coding, they were fleshed out in terms of their given properties and dimensions (Strauss & Corbin, 1998, p. 101). Axial coding aimed at developing a conditional matrix. To this end, categories were related to their subcategories and categories were linked to their properties and dimensions. In the final stage of data analysis the core category, in this case, **roots of concern**, was selected and systematically related to other categories. To establish trustworthiness, the provisional concepts and categories as well as the final version were confirmed and corroborated by the participants.

Despite the participants' validation of the emerged concepts and categories and methodological rigor, however, findings such as these are not to be taken as a guarantee of truth, for truths are always partial (Clifford, 1986), and knowledge "situated" (Haraway, 1988). We also cannot ignore how interviewer and interviewee negotiate face or manage impressions (Goffman, 1959) in interviews. An interview is but a snapshot in time. Much is left unsaid about events and persons despite the

intention of the interviewer to provide a holistic account. Of course, more interviews and stories would deepen our understanding of this exploratory study.

## Limitations of the Study

Although the researcher tried to validate final concepts and categories through member checking, findings such as these are not a guarantee of truth, for truths are always partial (Clifford, 1986), and knowledge "situated" (Haraway, 1988). We also cannot ignore how interviewer and interviewee negotiate face or manage impressions (Goffman, 1959) in interviews. An interview is but a snapshot in time. Much is left unsaid about events and persons despite the intention of the interviewer to provide a holistic account. Of course, more interviews in other contexts would deepen our understanding of this exploratory study.

## Results

### Roots of Concern

Description is the first step towards understanding a phenomenon; in this case the phenomenon being complaint and dissatisfaction pertaining to teacher evaluation. Description is limited to the effects, i.e., the visible. The second step is to explain the phenomenon by uncovering the causes, or the invisible. Thus understanding evolves by connecting visible effects to invisible causes. Thus the question is, "What are the roots of concern in the evaluation system of SUT?" Iterative data collection and analysis uncovered four main causes of concern: students' erroneous views, faulty evaluation tool, faulty administration of the tool, and limiting and limited decisions. In what follows, the study corroborates and validates these findings by relating them to extracts from interview data from the participants.

### Students' Erroneous Views

Students' views are usually erroneous in that they reflect many factors other than teaching performance. Scholars call into question student ratings of instruction since they have their doubts with regard to the validity of students' perceptions of teaching (Spoudle, 2002), and consider student rating as "meaningless quantification" and leading to "personality contests" (see Kulik, 2001), instead of measuring teaching performance. In this context, students' evaluation of teaching performance has caused lots of complaints among practitioners. But the main complaint is that evaluation puts the learning responsibility on teachers. Participants believe that students may fail to learn for many reasons that are unrelated to the act of teaching. Due to the nature of the university entrance exam in Iran, a great majority of students study something that they would not if they were free to choose. This leads to many other problems. For instance, there are students who study a discipline for which they do not have the pre-requisite knowledge. One participant complains:

> More often than not students of pure mathematics do not have the pre-requisite knowledge to study in this major. Due to the dominant social attitude and demand, bright students choose engineering as the first priority. The entrance exam divides students into the high ability and low ability groups; those who are not accepted in engineering, study

> mathematics. Not having the needed background knowledge, they cannot follow the instruction and as such evaluate the professor negatively.

Sometimes students evaluate instruction negatively because the class is totally heterogamous. Since there is no placement test prior to instruction, low ability and high ability students study in the same class. This factor negatively affects both learning and teaching. The following comments better explain the situation:

> In general English there are some students who enjoy a good command of English either because they have been in an English speaking country or because they have studied English in private institutes. On the other hand, there are a large number of students with a very low command of English because their studies were limited to high school syllabus. No matter how I teach, one group complains. If I respond to the demands of the high ability group, the low ability group complains because they cannot follow the instruction. On the other hand, if I respond to the low ability group, the high ability group complains because the class is very boring. Although heterogeneity is beyond my control, it negatively affects students' evaluation of my performance.

Another problem with students' evaluation of instruction is that a great majority of students in Iranian universities evaluate success in terms of their score in the final exam, rather than in terms of learning. Thus they evaluate teachers in terms of the item difficulty of the final exams and teachers' strictness in scoring rather than by the quality of instruction. When students aim at learning, they evaluate teachers' instruction. But when they aim at passing, they evaluate teachers in terms of their pass rate in the finals. The following comments are exemplary in the interviews:

> The problem with my students is that from the very beginning of the term they plan for passing rather than plan for learning. When I teach, their main question is, "Will you test the point you are teaching? If the answer is yes, they listen and make notes. On the other hand, when you answer an occasional question which really improves students' learning, they do not listen, if it is not covered in the specified syllabus. Students evaluate teachers' testing rather than their teaching. Your evaluation score is high if you have a high pass rate. If your pass rate is low this term, your evaluation score will be low the next term. Thus students should be allowed to evaluate the teacher once on the same course. When they evaluate the second time, their views are biased.

The number of students in a class affects the quality of teaching and learning and consequently affects students' perception of teaching performance. In very large classes, there is no time for questions and answers. Students do not find a chance to participate in classroom activities. In such classes, teachers' main concern is classroom management rather than quality teaching. One participant explains:

> Students' evaluation of my performance in a large class is different from their evaluation of my performance in small class. This is natural since in a large class there is not time for interaction with the students.

> There is no time to receive feedback and adjust teaching to respond to their needs. In a class with sixty students, the only thing I can check is their presence through calling the roll. In such a class there is no time for checking students' performance. Thus I do not know my students' level of knowledge and skill. This is withheld until the final exam. The results of final exams are the only type of feedback they receive in such a crowded class.

The other problem with students is that they want more for less; that is, they prefer a short booklet which is a shortcut to the final exams rather than an elaborate and demanding syllabus. More specifically, they prefer an objective method of instruction that guarantees success in the final exams. Having set a university degree as their goal, they seek shortcuts by cutting corners. Some professors resist this temptation. Take the following example:

> Practitioners' promotion depends on students' ideas. Students are aware of this. They shape instruction by imposing their likes. If a student is thirsty, and the professor does not know, the problem is with the professor, but the problem is that they want less. They evaluate me negatively not because I do not teach properly but because I do not surrender to their likes. Just like a responsible father, a professor should give his students what they need rather than what they want. The child does not like to take medicine. But the father makes the child take it. Sometimes as professors we should make students study up to their potential though they may reject it.

Some professors criticize students' evaluation of teaching performance by relating it to students' lack of background in evaluation. Since they never evaluated their teachers during high school, they do not know how to do it upon entry to the university. As they acculturate themselves to the university norms and values, their decisions get more reliable. Take these comments:

> During high school they have no part in decision-making. Proctors discipline them to do as they are told. They never think of evaluating teachers. Suddenly, they enter the university and they are asked to evaluate teachers. They have never evaluated anything. Professors who teach seniors and juniors have higher evaluation scores because these students have got used to university culture and decision-making.

**Faulty Evaluation Tool**

Before using any measurement and evaluation tool, evaluators should make sure that the tool measures consistently, i.e., it is reliable, and that it measures what it intends to measure, i.e., it is valid. Participants complain that the tool used to evaluate instruction at SUT has not been tested for these criteria. That is, we use a tool with dubious reliability and validity indexes. Similarly, at the level of items, participants complain that some items do not discriminate between low and high performance in teaching. The most frequent complaint is that items are mostly general. Since the evaluation tool was designed to be used for the evaluation of teaching performance in the entire university curriculum, it does not cover items that specify quality

performance in a specific discipline such as teaching English or physical education. There are many inherent differences in the methodology and technology of teaching these two distinct disciplines. None of these differences, however, is captured in the present items of the evaluation tool. If these differences are ignored, evaluation cannot provide effective formative feedback. One participant explains:

> …first the evaluation tool should be improved. We need a specific form which is in line with the objectives of the course. The general form cannot measure many aspects of the objectives of a specific course. One general tool cannot be used to measure professor's performance in physical education and electrical engineering. The main problem of such an evaluation is that it cannot give appropriate feedback for the improvement of performance. To improve teaching performance, feedback should be specific.

Others go beyond having a specific tool, which is used in parallel with the general tool. They suggest that students do not have the required knowledge and skills to evaluate technical aspects of teaching such as the adequacy of the syllabus, teaching methodology, and teachers' professional knowledge and skills. Thus the specific tool should be developed and administered by a professional committee in each department. One participant explains necessity of the specific form as follows:

> The present evaluation system does not and cannot evaluate the professors' professional knowledge, the syllabus and the degree to which they cover the syllabus. These aspects should be evaluated by a professional community. Before the term the professors should submit their syllabus and lesson plan to the committee. They should decide whether the syllabus and lesson plan are in line with the objectives of the course or not. Moreover, the committee should evaluate the professors' final and mid-term exams to find out their degree of compatibility with the objectives.

The second problem is that items have not been operationalized. As such participants describe items as ambiguous, two-dimensional, subjective and interpretable. One participant explains and exemplifies two-dimensional items as follows:

> I believe there should be more items. To avoid a lengthy questionnaire, they have put many criteria in one item. Take the item, "S/he can control and manage the class". This item contains two contradictory criteria: “control” carries a negative connotation while 'management' carries a positive connotation. Whatever it is, they are not the same thing; however, they are measured in one item. I do not agree with the concept “control” at all. To control the class, it takes a dictator. To this end, s/he can act as a commander and s/he may see students as soldiers. Thus s/he can control the class, but is this pedagogically acceptable?

Moreover some items are open to the subjective interpretation of students. Different students interpret them in different ways since they do not carry a single

objective meaning. One of the participants describes the item, "S/he observes professional etiquettes" as being totally subjective. He explains:

> This item is multi-dimensional. The “s” in the word “etiquettes” clearly indicates that the item measures many different things. But it does not specify what these things are. For each student it may denote one or more specific meanings. Thus students interpret it in quite different ways: for one it may denote “verbal behavior”, for anther it may denote “clothing” and still for another it may denote “order or discipline”. Anyhow, I myself cannot get what the item means by the concepts “professional” and “etiquette” since they have never been specified and publicized in advance.

Some items are rejected on the ground that they are dependent. For instance transmission of content (item one) depends on mastery over content (item five). Similarly transmission depends on methodology and technology (item two). Item two is two-dimensional. It measures methodology and technology at the same time. It is methodology that determines technology. Thus technology depends on methodology. Item dependency creates areas of overlap. One of the participants explains:

> I believe there is eighty percent overlap between items one and two. The use of suitable methodology and technology greatly facilitates transmission of content. On the other hand, transmission of content proves very difficult without the use of appropriate methodology. I believe everything depends on methodology. A method clearly specifies the role of technology, the syllabus, the teacher and the students.

But participants' main concern is that the evaluation tool oversimplifies the distinctive features of teaching performance in only two items. Methodology and syllabus design are holistically measured by items two and three. To evaluate the syllabus alone, there are standard and validated tools containing more than ten items. The oversimplification of the core of classroom activities in two items has led to the faulty and simplistic weighing of items. In its present form, the adequacy of the syllabus carries the same weight as discipline. One participant explains the consequences as follows:

> Take two teachers: one observes professional etiquette but his syllabus it totally outdated, inefficient and inappropriate while the other does not observe professional etiquette, something that is open to students' subjectivity since it is not well-specified and defined, but his syllabus is up-to-date, efficient, and relevant to students' needs. Now suppose that students evaluate the former as very week in relation to the syllabus and evaluate the latter negatively in terms of his professional etiquette. Everything else being equal, these professors will have the same performance score. This decision is far from fair. Although professional etiquette is important, its instances should be specified. Moreover, it should carry far less weight than the syllabus.

**Faulty Administration**

A well-developed and comprehensive evaluation tool that is both reliable and valid can yield faulty results if it is not administered under uniform and standard conditions. Bad administration can produce error variance. Participants are not satisfied with the time of administration, confidentiality of students' responses as well as confidentiality of evaluation records. They believe that evaluation is not fair if the effects of these factors are not controlled. One of the most frequent complaints was related to the administration of the evaluation tool nearly at the end of term. One participant explains the problem as follows:

> There are two groups of students in each class. Those who have been present during the term may leave the class towards the end of the term to prepare themselves for the finals. On the other hand, there are some students who are guilty of absenteeism during the term. To compensate for their absence they participate toward the end of the term. Since evaluation tool is administered nearly at the end of the term, mostly those who have been absent beyond limit are present in evaluate the professor. But the main problem is that a great number of students are absent. In a forty eight-student class only ten students evaluated my performance. In another class only five students evaluated the course. Can we take the views of a limited number of students who have been mostly absent during the term as a representative sample of the class population as a whole?

Another problem is that students do not fill out the tool under uniform conditions. The presence of professor in the evaluation session can greatly shape students' responses. The problem is that some professors leave the class when the tool is administered while some others stay in the classroom. This not only creates a high level of anxiety among students but also shapes their responses. Participants' comments better explain the situation:

> The administration officer told me that some professors stay in the class and monitor students' responses. With their presence the professors make the students evaluate them the way they like. On the other hand, when the professor leaves students can freely express their views. I believe that the evaluation score of a professor who does not leave the class reflects censored views rather than students' real views. Administration should guarantee the confidentiality of students' views.

The third problem with administration is that evaluation records are exposed to the views of many outsiders since they are not signed and sealed immediately after the administration. This may create problems for both students and professors. Students respond on the ground that their views remain confidential. Similarly, professors want their evaluation record to be in safe hands. Participants complain that confidentiality of records is not observed. One participant complains:

> The evaluation officer comes and collects evaluation data. Then he keeps them in his office for one or more weeks. During this time the records may be manipulated by interested parties. Similarly, they may

> be viewed by various groups including the professors themselves. If the professor observes students' views, he may use negative views against them. Thus I believe that evaluation records should be and signed and sealed the very moment they are collected. Data should remain confidential so that professors and students trust the evaluation system. We should clarify the parties that should have access to the data, and state the reason for their access.

The final factor that can affect evaluation results is the class time. Some classes are held in the morning while other classes are held very late in the afternoon, e.g., six to eight in the afternoon. This time span is a time when both the students and professors are exhausted. One of the participants explains:

> Physical strength and mental concentration greatly affects teaching and learning performance. As a rule both professors and students are exhausted at this time of the day. Some professors may plan to control this negative effect. But can they control this effect in his students? In such a class the professor may perform well. Since the students are exhausted, however, they cannot follow the instruction. Such a class is not efficient in terms of student achievement. Low achievement negatively affects students' evaluation of teaching performance.

Participants suggest that evaluation improves if the tool is administered under standard and uniform conditions. As such evaluators should specify a set of guidelines for evaluation officers. Even some suggested that evaluators should be trained. Quality evaluation minimizes the effect of unwanted sources of variance such as the ones introduced by faulty administration. Teachers' evaluation score should reflect their teaching performance and nothing else. This is possible if evaluators systematically control any other sources of variance.

**Limiting Decisions**

Due to the nature of the course and many other factors, students' evaluation of teaching performance is high in some courses and departments but it is low in some other courses and departments. Thus comparing teaching performance in one course from one department with another course from another department is very misleading. One participant complains:

> In "General English" my performance score was 3.5 and in "Communication Theories" it was 3.0. On the surface, it is clear that I performed better in general English. But if you compare them with the group means in the two departments, you will have another decision. Although my performance score seems to be lower in "Communicative Theories", I performed well in the related department since my performance is well above the group mean. On the other hand, when compared with group mean, my performance in general English is not good because my evaluation score is lower than the group mean. To compute my mean score, the evaluators added these two up. It is totally illogical; it is as illogical as 3 apples + 2 bananas.

The method of interpreting evaluation scores is similarly limiting. To interpret scores, you can either compare the raw score with the mean of the group, i.e. norm-referenced evaluation, or compare each individual score with a previously established criterion such as an accepted level of mastery, i.e., criterion-referenced evaluation. The second approach has the potential to give the evaluator a list of those who do not have an accepted level of mastery. The evaluator can use this information to plan workshops for teacher development. The first approach is limiting in that it does not clarify level of mastery; it only gives the position of each individual in the group. One participant explains:

> Evaluation identifies low performance and high performance but it never improves teaching performance, and low performing professors stay the same. It describes the status quo and preserves it because it does not specify the consequences for low performance and high performance. Its negative side-effect, however, is quite evident: it creates jealousy and suspicion among colleagues. Those who perform well are suspicious of inflating students' pass rate in the finals. When you perform well, colleagues believe that students' ideas are not reliable. When your score is low, however, it is taken as a hard and fast rule.

Decisions made on the basis of performance scores are limited in two ways. First, they are limited because their source of data is limited, i.e., decisions are based solely on students' evaluation of teaching performance. As stated in the review of related literature, information can be and should be collected from multiple sources. Decisions which are based on multiple sources of data are inherently more rigorous and reliable than decisions which are based on a single source of data. This is what participants are well aware of. One participant explains:

> We should not promote the professors or blame them for incompetency solely on the basis of students' views. I do not say that students' views are wrong. But I do believe that they are partial. To get a better picture, we should evaluate teachers from at least three sources: students' views, colleagues' views in this department, and colleagues' views at the university. Then we should weigh these sources properly.

Another participant goes further by stating that students cannot evaluate professors' professional knowledge and skills or the adequacy of the course syllabus. These are technical aspects, which should be evaluated by a professional committee in the department. Students can evaluate general aspects of teaching as done currently. As for the technical aspects he states:

> … students' evaluation of teaching performance is very superficial. Their evaluation reflects the professor's relationship with the students, his sociability, his temperament, his strictness in scoring and counting the presence of students. They can only evaluate the degree to which a professor's teaching is comprehensible. The present evaluation system does not and cannot evaluate the professor's professional knowledge, the syllabus and the degree to which he covers the syllabus. These aspects should be evaluated by a professional community. Before the

> term the professor should submit his syllabus and lesson plan to the committee. They should decide whether the syllabus and lesson plan are in line with the objectives of the course or not. Moreover, the committee should evaluate the professors' final and mid-term exams to find out their degree of compatibility with the objectives. We cannot leave these technical issues to students' judgment.

Another limitation of the current evaluation system is that it does not provide constructive feedback that could be used to reform educational ills. Focusing on the general inter-departmental aspects of teaching, the evaluation system leaves technical aspects to chance. The feedback such a system provides is general. To reform instruction, feedback should be specific. Participants' comments better explains the problem:

> In general aspects of teaching I can use students' views to improve teaching. For instance, last term students complained that I was not regularly present in the specified consulting hours. This term I tried to solve this problem. But when students say that they are not satisfied with my teaching methodology, I do not know what I should do to improve instruction. I do like to improve my teaching. To do that, however, I need specific feedback on specific aspects of teaching. I do not accept that my teaching methodology is ineffective as a whole. But I do accept it if it specifies the aspects I do not perform well. Evaluators cannot evaluate my teaching methodology with one general item. The tool needs specific items that cover different aspects of teaching.

But the main limitation of the current evaluation policy is that it does not have any effect on the quality of teaching and learning. Thus it is limited because it is used mainly to give tenure, promote and dismiss. But the evaluator himself acknowledges that the current evaluation system fails to fulfill even these functions. He states:

> We use the results for promotion. But there are some who become indifferent after promotion. We also use it to give tenure. But the problem is that after receiving tenure, the evaluation score of some professors drop. Interviews with students show that some of these professors are very bad-tempered, they are not punctual. Evaluation is also used to control bad performance. If a professor's evaluation score is below 2.5 for two consecutive terms, his teaching hours are minimized. But the problem is that in some departments, especially in some courses, we do not have enough professors. This strategy fails in such cases. At the time being, there is no reward for those who perform well. In future we are going to use standard evaluation scores to reward those who perform well.

Thus our evaluation system is limited in that it ignores formative evaluation and focuses solely on summative evaluation. But as the comments of the evaluator indicate, the policy fails even in its summative functions, i.e., in promoting, giving tenure, and dismissing. Summative evaluation does not aim at improving instruction.

To this end, the system should introduce formative evaluation. It is this function that participants unanimously agreed and suggested. One participant explains:

> We should not use evaluation to compare one professor with another. We are much better off if we use evaluation to diagnose weaknesses and strengths in teaching performance. Based on the results of diagnostic evaluation, we can plan workshops to address and improve weaknesses and recognize, reward and publicize strengths. But the present tool is not good for formative evaluation. The tool should be developed in a way that it gives specific feedback. The current tool is too general to give formative information.

**Consequences**

As discussed in the results section, one major limitation of the evaluation system at SUT is that it focuses on summative evaluation. This approach has two major negative consequences. First, it focuses on poor teacher performance and gathers documentation on a teacher's weaknesses. Hence being judgmental, it has created an atmosphere of suspicion among practitioners. Second, it does not recognize and reward merit performance. Thus it has created a state of indifference. More specifically, when students' evaluation of teaching performance identifies the low group, they are taken as valid. But when they identify merit performance, they are taken as unreliable and invalid. Low performing group accuse the high performing group of inflating students' final scores and being lax in controlling students' presence in the classroom. One participant explains:

> When I received my evaluation report from the department and the head of the department saw my evaluation score, he said that students favor me because of high pass rates in your final exams. He further accused me of being friends with the students and being lax in classroom control. He complained that his evaluation score is low because he is strict in calling the rolls and scoring the final exams.

Moreover, since evaluation does not involve any incentive scheme, it has created a state of indifference among both the low performing group and high performing group. Similarly students do not take evaluation seriously. One participant explains:

> Nearly everyone is indifferent. The reason is that there is no difference between those who score low and those who score high. They are indifferent because they do not see the effect of evaluation in the environment. We should use the result of evaluation to improve the performance of those who scored low. We should hold teaching workshops for them. Since there is no room for improvement, the low performing group is indifferent. Those who scored high should be positively reinforced; their performance should be recognized and publicized. They should act as role models. Since this has never happened, this group is indifferent too. They use the results only for promotion. The problem is that some professors do not take promotion seriously. They know that they cannot promote because they do not

> have the required research credits. Students are similarly indifferent. They see professors with very low evaluation scores teaching the same subject again and again without any improvement. They believe that evaluating professors is a waste of time.

Practitioners relate indifference to a lack of clear policies for high and low performance. On the other hand, those in charge of evaluation system relate it to practitioners' being irresponsible. Take the following comments from the evaluation and supervision office:

> The evaluation tool is not comprehensive. We should have specific items for each department. We asked the professors of all the departments to send us specific items. We received only one response. The others did not respond at all. Some professors do not even consult their evaluation report. They are irresponsible.

But the most limiting aspect of the evaluation system is that by ignoring formative evaluation, it has left no room for professional growth. Almost all the professors teaching in this university have no systematic background in teaching methodology and testing since they graduated from universities of technologies. Thus they are in urgent need of methodological innovations. By focusing on the summative function of evaluation, the system does not candidly respond to this urgent need. Evaluation scans poor performance but it does not candidly address it. One participant explains:

> Since students' evaluation of my teaching performance was low last term, this term my department minimized the number of credit units I can teach. I have two objections to this decision: first they negatively judged my professional knowledge and skills in its totality based on students' views; second, they did not specify areas of weakness. Evaluation should not penalize colleagues for poor performance. It should diagnose areas of weakness and then systematically plan to obviate them through workshops and teacher development groups. They are many others like me who continue teaching without knowing or improving specific areas of weaknesses.

## Discussion and Conclusion

If you compare the rhetoric of teacher evaluation, as presented in the introduction section, and the practice of teacher evaluation at SUT, as discussed in the result section, you will find a wide gulf between theory and practice. Promising theories and bleak practice indicate that in this locality and in many similar contexts, teacher evaluation follows local traditions rather than research findings. By systematically juxtaposing theory and practice, we can identify two major shortcomings: reliance on a single source of data and adherence to summative evaluation. To address and obviate these problems, the evaluation system of SUT should:

1. Base its decisions on multiple sources of data such as peer review of teaching, review of lesson plans, classroom observation, and

portfolio review rather than limit itself to a single source of data, i.e., students' evaluation of teaching performance.
2. Make use of both summative and formative evaluation rather than limit itself to the summative function of evaluation that shuts the door to any improvement. When coupled, formative and summative evaluations provide optimal professional development opportunities (see Nolan & Hoover, 2005) and tenure (Brandt et al., 2007).

The evaluation system can systematically address and obviate these shortcomings by accommodating the fore-mentioned research findings and theoretical perspectives. It can similarly reform the current trend in evaluation by accommodating the findings of this study. Accommodating the views of the participants in this study entails teacher evaluation by teachers for teachers. To improve students' evaluation of teaching performance, the evaluation system should:

1. Empirically establish the reliability and validity of the evaluation tool;
2. Empirically establish item discrimination, i.e., provide empirical evidence that the items systematically discriminate between high performing group and low performing group;
3. Statistically convert raw scores to standard scores and then compare performance based on standard scores rather than raw scores;
4. Intra-departmentally develop and administer a specific tool to be used in a parallel fashion with the general tool;
5. Empirically separate the variance related to contextual and learner constrains from the variance related to teacher constraints. At the time being evaluation puts the full responsibility of learning on teaching. Logically, teachers are not responsible for learner and contextual constraints;
6. Rigorously correct items for subjectivity, conditionality, dependency, overlap, relevance, practicability, and ambiguity;
7. Rigorously minimize the effect of faulty administration by administering the tool under uniform conditions;
8. Systematically compare individual teaching performance with previously established performance criteria rather than compare the teaching performance of professors of electrical engineering with that of physical education; and
9. Systematically use evaluation as a scientific mechanism for creating conditions that are conducive to professional development rather than suspicion and indifference.

The significance of the findings is manifold. First, they are significant for practitioners themselves since through the dialogical process of grounded theory, practitioners realize how evaluation criteria shape their performance. Second, they are significant for local evaluators since they are provided with a rich source of empirical data grounded in practitioners' perspectives for reform. Such a bottom-up reform enhances job satisfaction among practitioners since they provide them with an evaluation system which is grounded in their own views. Third, they provide policy

makers at the Ministry of Science, Research and Technology with a rich source of data for national improvement of evaluation policy since students' evaluation of teaching performance is not limited to the research context. Finally they are significant for other countries following similar strategies by presenting them with consequences of evaluation policy on practitioners' professional life.

## References


Attinello, J. R., Lare, D. W., & Source, F. (2006). The value of teacher portfolios for evaluation and growth. *NASSP Bulletin, 90*(2), 132-152.

Berry, B. (2007, March). *Connecting teacher and student data: Benefits, challenges, and lessons learned.* Presentation for the Data Quality Issue meeting, Washington, DC. Retrieved February 25, 2008, from http://www.dataqualitycampaign.org/files/Meetings-DQC_Quarterly_Flyer_031207.pdf

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). *Examining district guidance to schools on teacher evaluation policies in the Midwest Region* (Issues & Answers Report, REL, 2007-No.030). Washington, DC: U.S. Department of Education, institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved February 25, 2008, from http://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_2007030_sum.pdf

Brown, S. C. (1999). *Learning across the campus: How college facilitates the development of wisdom.* Unpublished doctoral dissertation, University of Maryland, College Park.

Clifford, J. (1986). Introduction: Partial truths. In J. Clifford & G. Marcus (Eds.), *Writing culture: The poetics and politics of ethnography* (pp. 1-26). Berkeley, CA: University of California Press.

Crew, L. A., Everitt, T. J., & Nunez, R. W. (1984). *Improving teacher performance through systematic teacher evaluation.* Virginia Beach, VA: American Association of School Personnel Administrators.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional learning.* Princeton, NJ: Educational Testing Service.

Glaser, B. (1978). *Theoretical sensitivity.* Mill Valley, CA: Sociology Press.

Goffman, E. (1959). *The presentation of self in everyday life.* New York, NY: Doubleday Anchor Books.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies, 14*(3), 575-599.

Kulik, J. A. (2001). Student ratings: Validity, utility and controversy. *New Directions for Institutional Research, 27*(5), 9-25.

Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan, 86*(10), 727-735.

Mujis, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53-74.

Nolan, J., Jr., & Hoover, L. A. (2005). *Teacher supervision and evaluation: Theory into practice.* New York, NY: Wiley.

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousand Oaks, CA: Corwin Press, Inc. (ERIC Document Reproduction Service No. ED445087)

Peterson, P. L., & Comeaux, M. A. (1990). Evaluating the systems: Teachers' perspectives on teacher evaluation. *Educational Evaluation and Policy Analysis*, *12*(1), 3-24.

Price, J., & Schwabacher, S. (1993). *The multiple forms of evidence study: Assessing reading through student work samples, teacher observations and tests.* New York, NY: National Center for Restructuring Education, Schools and Teaching.

Seidman, I. E. (1991). *Interviewing as qualitative research: A guide for researchers in education and the social sciences.* New York, NY: Teachers College Press.

Shannon, D. M. (1991, February). *Teacher evaluation: A functional approach.* Paper presented at the annual meeting of the Eastern Educational Research Association, Boston, MA.

Sproudle, R. (2002). The un-determination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review, 21,* 287-294.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

Stronge, J. H. (2007). Planning and organizing for instruction. In J. H. Stronge (Ed.), *Qualities of effective teachers* (pp. 212-243). Alexanderia, VA: Association for Supervision and Curriculum Development.

Tucker, P. D., Stronge, J. H., & Gareis, C. R. (2002). *Handbook on teacher portfolios for evaluation and professional development.* Larchmont, NY: Eye on Education.

Uhlenbeck, A. M., Verloop, N., & Beijaard, D. (2002). Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*(2), 242-272.

---

## Author's Note

Seyyed Ali Ostovar Namaghi was born in 1969. He received his MATEFL from the University of Tehran and then entered Shiraz University for the PhD programme. He achieved rank 1 in the Comprehensive Exam of Shiraz University for PhDTEFL candidates. His chief research interest is language teacher knowledge. He has published in a number of leading peer-reviewed journals including: The Reading Matrix, Teacher Education Quarterly, and Asian EFL Journal. Presently he runs courses in EAP at Shahrood University of Technology. Correspondences regarding this article can be addressed to: Seyyed Ali Ostovar Namaghi, PhD; Department of General Courses, Shahrood University of Technology, Shahrood, Iran; Phone: (98)9177110710; E-mail: namaghisa@yahoo.com

This study was funded by the Research Department of Shahrood University of Technology. The researcher wants to thank Dr Pooyan and his colleagues at this department, whose assistance was invaluable in conducting this study.

**Article Citation**

Ostovar Namaghi, S. A. (2010). A data-driven conceptualization of teacher evaluation. *The Qualitative Report, 15*(6), 1504-1522. Retrieved from http://www.nova.edu/sss/QR/QR15-6/ostovarnamaghi.pdf