Rater Effect in 360-Degree Teacher Evaluation: A Case of Teaching Effectiveness in Thailand*

Jittima Juntavech

Ratchaneekool Pinyopanuwat

Sungworn Ngudgratoke

ABSTRACT

The purpose of this study was to investigate the rater effect and the halo effect in the 360-degree teacher evaluation. Samples consisting of 97 teachers, 186 peers and 1,912 students were drawn from school districts in Bangkok. The teaching effectiveness scale consisting of 47 items developed by Juntavech (1999) was used as the evaluation tool to evaluate rater and halo effect. Analysis of variance (ANOVA) and confirmatory factor analysis (CFA) indicated that students evaluated their teacher in a way that differs substantially from teacher and peers did. Another result suggested that halo effect exists in 360-degree teacher evaluation. The implication of this research for practices was that teacher evaluation should not rely on scores from a single evaluator and that 360-degree teacher evaluation might be appropriately used for teacher's continuous improvement but not for teacher promotion, hiring, or other administration decisions.

* Paper presented in the Annual Meeting of the American Educational Research Association, San Francisco, April 7-11, 2006.

Introduction to Teacher Evaluation

Over the decades, educators, researchers, and parents have emphasized on enhancing and maintaining teacher quality (Johnson, 1997; Wright, Horn, & Sander, 1997). Many researchers investigated the teacher factors that empirically influence student learning in order to seek better practices for upgrading teacher quality (Kaplan, & Elliott, 1997; Wright, Horn, & Sander, 1997; Birenbaum, 2003). The heart of this line of research is the core belief that teachers make a difference in student performance. To enhance teacher quality will ultimately improve students' performance.

The results of the research carried out to evaluate teacher effectiveness show very consistent findings across studies. That is, teacher factors explained such a large variation in students' performance (e.g., mathematics achievement). For instance, the quality of classroom instruction and classroom management which indicate how well a teacher performs in classroom were found to be important factors to enhance school learning (Wang, Haertel, & Walberg, 1993), and also teacher effects were dominant factors affecting students academic achievement gain (Wright, Horn, & Sanders, 1997).

With the emphasis on maintaining and enhancing teaching effectiveness, there has been an increasingly interest in teacher evaluation across nations. Teacher evaluation has been expedited by two reasons. First, teacher evaluation has been conducted because it serves as a diagnosis and reflective tool for improving teacher professional development and promoting pedagogical reform (Chen, Burry-Stock, & Rovegno, 2000). Second, teacher evaluation is also fostered by educational reform and educational accountability. The core idea of these reforms is to assure that students would receive better instruction and that a teacher will increase their instructional productivity which ultimately improves the quality of educational system. In other words, teachers are to be evaluated in terms of their skills and competencies needed for delivering good instruction. Stronge (1997) recommended that a conceptually sound and properly implemented evaluation system for teachers is a vital component of any successful reform effort and a vital part of total improvement.

Teacher Evaluation in Thailand

In the past teacher evaluation in Thailand was mainly used for teacher promotion in which the degree to which a teacher achieved the acceptable level of teaching quality was viewed by external evaluators. This blinds the understanding of the relation between teacher quality and student's learning and of teaching and learning activity that actually happen in the classroom.

Currently, there are major changes in teacher evaluation in Thailand. Teacher evaluation has been forced by the new educational legislation lanced in 1999. It has been stated in the legislation that evaluating teacher quality is a vital component of the educational quality assurance system. In addition, teacher evaluation is a tool for diagnosing the healthy of teacher development implemented within the teacher quality assurance system.

There is an agreement among Thai educators and researchers on that to evaluate teacher effectiveness it is better to include a variety of assessment methods such as peer evaluation, self-assessment, portfolio assessment. However, there has been very little if any work done on how to design and implement methods of teacher assessments that are congruent with the actual requirements in the new educational legislation (Pillay, 2002).

360-Degree Performance Assessment

360-degree performance assessment is characterized by the evaluation of an individual performance by multiple raters from multiple levels (Mount et al, 1998). This concept in teacher evaluation context closely parallels the movement to seek continuous performance improvement by using multiple data source to give feedback in teacher evaluation. Mount et al recommended that multiple raters are used frequently to enhance personal development and growth, rather than to help with salary administration, promotions, or other administration decisions. Stronge (1997) supported that the use of multiple data sources for documenting performance is an important feature of an effective teacher evaluation system and that client feedback is the major source for teacher evaluation.

The two important reasons to use multiple raters were suggested by Mount et al. First, job performance is multidimensional; thus, multiple raters may be better suited to evaluating certain aspects performance. This implies that the use of multiple data sources to measure teaching effectiveness can result in a fuller and more accurate view than is available through more narrowly defined approach to data collection. Second, raters bring different views and aspects to evaluate teachers. Even thought they have the same opportunity to observe performance, they may perceive and evaluate it differently.

A major concern of using the 360-teacher evaluation was recommended by Wilson (1988) that we should use multiple assessors who are familiar with the skills for teaching needed to be assessed. Millman (1981) pointed out that even though many authors and researchers recommended using multiple raters to assess teacher performance, all raters are not equally suited to evaluate all aspects of teaching. Teachers' self-evaluation, student evaluation, and peer assessment are only key raters of teacher evaluation (Elmier, Jenkins, & Crawford, 1991; Stufflebeam & Shinfield, 1985; Airasian & Gulickson, 1997; Davis, Ellett, & Annunziata; 2002).

Even though the 360-teacher evaluation is fruitful, controversial issues of using multiple raters in teacher evaluation have been highlighted. First, one might ask if it is possible to use students as evaluators of teaching effectiveness. Even though many researchers such as Jackson (1999) suggest that multiple evaluators could be used to fuller understand of multidimensional nature of teaching effectiveness, there is still a disagreement on its uses. Second, it is not guaranteed that using multiple evaluators to evaluate teacher's performance is free from rater effect (i.e., halo effect). The halo effect generally occurs when one positive factor overshadows all negative factors and produces an artificially high summary score (Shepard, 2005). Thus, the valid outcomes of teacher evaluation using multiple evaluators might be concealed if rater effect exists.

Purpose of the study

The purpose of this study was to evaluate if and what extent to which rater effect exists in 360-degree teacher evaluation. Rater effect in the present study was characterized by the inconsistency of ratings given by different types of raters. Also, halo effect in 360-degree performance assessment was investigated. The findings would be used to seek implications for decision-makings concerning the uses of evaluation results.

Methods

Sample

There were three different samples in this study which were teachers, peers, and students. The samples consisting of 97 teachers, 186 peers and 1,912 high school students in the schools located in Bangkok, Thailand, were selected. The teacher evaluation data were collected in 1999. Specifically, 38 schools in Bangkok were randomly selected and at most 3 teachers were chosen from each school. Then, two or three peers of a chosen teacher were randomly selected to be evaluators responsible for providing peer evaluation results. Finally, one classroom in which students were taught by a chosen teacher was selected and every student in the chosen classroom was asked to evaluate their teacher's performance in teaching.

Instrument

Data used in this study were the secondary data collected by Juntavech (1999). Originally, the 47 items intended to measure teaching effectiveness were developed based on the 6 domains: teacher's characteristics, subject matter knowledge, attitude toward students, interpersonal interaction, instructional competency, and assessment competency. This instrument was the rating scale consisting of 6 ratings ranging from 1 to 6, where 1 implies a teacher had not perform an activity being asked by the item, and 6 implies that a teacher had performed the activity being asked by the item.

Experts including an experienced teacher, a professor, an educator, and a measurement expert were included in the process of gathering evidences for judging the degree of content validity. After the content validity was supported by experts' agreement, the instrument was piloted with 2 teachers, 4 peers, and 40 students. These samples were not the same as those in the final data collection stage. Cronbach's alpha reliability coefficients were .95, .96, and .94 for the group of teachers, students, and peers, respectively.

Note that the evaluation instruments for teachers, peers, and students were actually the same in terms of content domains being evaluated and number of items, except for the part I which was related to demographic information.

Data Analysis

The data analyses proceeded as follows:

1. Descriptive statistics such as frequency and percent were analyzed to show the demographic characteristics of teachers, peers, and students who participated in the present study.

2. Exploratory factor analysis (EFA) was used to explore the number of dimensions existing in the teacher evaluation data evaluated by teachers, peers, and students. The dimensions which were extracted by factor analysis using principal factor extraction method then were used as the performance dimensions used to evaluate teachers' teaching effectiveness. In other words, in stead of comparing the scores on individual items evaluated by teacher, peers, and students, the extracted performance dimensions obtained from exploratory factor analysis would then be used as the scores used to evaluate teachers' teaching effectiveness.

3. The rater effect was determined in the present study as degree of the inconsistency in evaluation data evaluated by three different types of raters: student, peer, and teacher. The analysis was performed for both item and performance dimensions. To evaluate whether the rating results from different types of evaluators were consistent, analysis of variance (ANOVA) was used to evaluate if rater effect exists. If the mean ratings from at least two sources were statistically significant, this would indicate the existence of rater effect.

4. Multitrait multimethod (MTMM) which was conducted through confirmatory factor analysis (CFA) was used to identify whether halo effect existed in 360-degree teacher evaluation. CFA offers a more systematic way to analyze multitriat-multimethod data than simple inspection of correlations (Kline, 1998). Studies that have examined performance rating data using MTMM usually focus on the proportion of variance in performance rating that is attributable to traits and that which is attributable to raters. In most MTMM context, it is desirable to have a high proportion of trait variance and low proportion of rater variance. The predominance of rater variance over trait variance suggests that the existence of rater effect or halo effect which occurs when evaluator's ratings are heavily influenced by an overall evaluation of the ratee.

CFA model assumes that each evaluation dimension contains rater variance, trait variance, and unique variance. That is, teaching effectiveness and raters were considered latent variables or factors but the teaching effectiveness was assume to be independent of rater factors, while the extracted dimensions obtained from EFA were the observed variables. This analysis enables researches to explore if, in addition to the teaching effectiveness, rater traits had impact on teacher evaluation results. Halo effect was measured by comparing factor loadings that loaded on both raters and teaching effectiveness dimensions.

To evaluate the fit of the MTMM model, six indices were used. These indices included the chi-square (χ^2) index, the goodness-of-fit index (GFI), the nonnormed fit index (NNFI), the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the root mean residual (RMR). These fit indices were chosen because no single fit index is considered to be the definitive marker of a model with "good fit"; each index serves a different purpose and should be interpreted in combination with the other indices. The χ^2 index is an absolute index that tests for lack of fit resulting from over identifying imposed on a model and sample size. Value of 1 for GFI and the NNFI indicates perfect model fit; however, some researchers have suggested cutoff values greater than .95 to indicate model fit. The following index cutoff values suggested by Hu and Bentler (1999) were used for determining good ness of fit: CFI > .95, RMSEA < .06, and RMR < .08.

Results

Descriptive statistics

Table 1, 2 and 3 present the demographic variables of teachers, students, and peers, respectively. As seen in table 1, the total number of teachers participating in the study was 97 teachers. Most of them were female teachers (84.54%). The majority of those teachers were 41 to 50 years old (75.29%). Moreover, most of them had bachelor degree (75.26%), more than 20 years of teaching experience (53.6%), were 10th grade teachers (38.14%), and had experience in teaching evaluation (79.38%).

Table 2 presents the characteristics of students participating in the present study. The majority of these students was female (64.8%), 16-17 years old (68.46%), 11th the grade students (39.85%), had grade point average ranging from 2.00-2.49 (30.13%), and most of them were those majoring in science and mathematics (53.14%).

Finally, table 3 presented peers' characteristics participating in the present study. Generally, the majority were female (79.03%), 41-50 years old (59.68%), and were 11th grade teachers (39.25%). They had bachelor degree (85.48%) and had experience in teaching evaluation.

Table 1. Teacher's Demographics (n=97)

Demographic variable	frequency	percent
1. Gender:		
Female	82	64.80
Male	15	15.46
2. Age:		
21-30	1	1.03
31-40	13	13.40
41-50	73	75.29
51-60	10	10.31
3. Education:		
lower than bachelor degree	2	2.06
bachelor degree	73	75.26
master degree	21	21.65
Ph.D.	1	1.03
4. Teaching experience (years):		
< 5	1	1.03
5-10	7	7.22
11-15	9	9.28
16-20	28	28.87
> 20	52	53.60
5. Grade level currently teaching:		
10th grade	37	38.14
11t grade	34	35.05
12th grade	26	26.81
6. Experience in teaching evaluation:		
Have experience	77	79.38
Have no experience	20	20.62

Table 2. Student's demographics (n=1,912)

Demographic variable	frequency	percent
1. Gender:		
Female	1,239	64.80
Male	673	35.20
2. Age:		
14	26	1.36
15	283	14.80
16	652	34.10
17	657	34.36
18	276	14.44
19	17	0.89
21	1	0.05
3. Grade:		
10th grade	585	30.60
11th grade	762	39.85
12th grade	565	29.55
4. Major:		
Science-Mathematics	1,016	53.14
Mathematics-Language	390	20.40
Social study-Language	130	6.80
Other	376	19.66
5. Grade point average:		
< 2.00	276	14.43
2.00-2.49	576	30.13
2.50-2.99	524	27.41
> 3.00	536	28.03

Table 3. Peers' demographics (n=186)

Demographic variable	frequency	percent
1. Gender:		
Female	147	79.03
Male	39	20.97
2. Age:		
21-30	10	5.38
31-40	25	13.44
41-50	111	59.68
51-60	40	21.50
3. Education:		
bachelor	159	85.48
master	26	13.98
Ph.D.	1	0.54
4. Teaching experience (years):		
less than 5	7	3.76
5-10	18	9.68
11-15	17	9.14
16-20	42	22.58
more than 20	102	54.84
5. Grade level currently teaching:		
10th grade	70	37.63
11th grade	73	39.25
12th grade	43	23.12
6. Experience in teaching evaluation:		
Have experience	129	69.35
Have no experience	57	30.65

In general, the teacher performance ratings evaluated by teachers were very similar to those evaluated by their peers. However, the results of teacher's performance evaluated by students were quite different from those evaluated by teachers and their peers. Figure 1 compares the differential distribution of mean ratings for each item evaluated by teachers, peers, and students. It can be seen that teachers tended to provide relatively high scores on every item being asked, while students tended to provide relatively low ratings. Peers were likely to provide the ratings which were in the middle between teacher's and students' ratings.

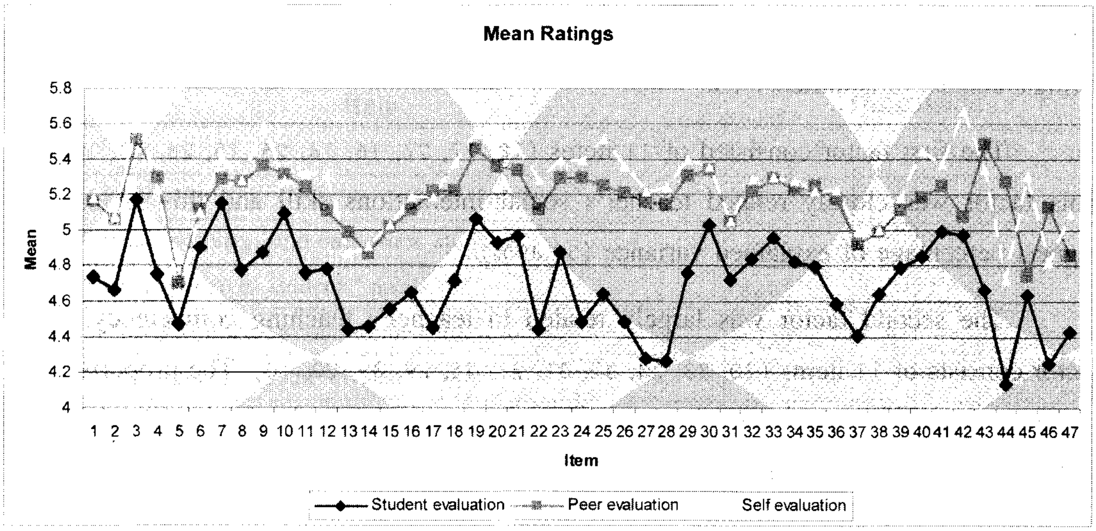

Figure 1. The Comparison of mean ratings for the individual 47 items

Analysis of variance

Analysis of variance was performed to explore if the ratings of the individual items item evaluated by three different types of raters were statistically significant. The results indicated that, for every item, *F*-tests are statistically significant at .05. Post hoc comparisons suggested that ratings given by teachers and peers were not statistically different. However, ratings given by students were significantly different from those given by both teachers and peers.

Exploratory Factor Analysis

In this section, factor analysis was used to determine the numbers of dimensions existing in the teacher evaluation data. The evaluation results from three sources were merged that resulted in 2,106 observations used in this analysis .The varimax rotation method was used to extract the numbers of dimensions in order to obtain more simple structure. For ease of interpretation, factor loadings which were less than .30 were suppressed.

The results of factor analysis indicated that there were 6 dimensions that had the eigen value greater than 1.0. Overall, there was 57.71 percent of the total variation in the evaluation data that was explained by the 6 extracted dimensions. In other words, it is reasonable to convince that the exploratory factor analysis showed that the teaching effectiveness data as evaluated by teacher's self-assessment, students' ratings, and peers' ratings, consist of 6 dimensions.

The first factor consisted of 11 items (18, 17, 27, 16, 14, 26, 15, 28, 13, 19, 9). This factor was closely related teacher's social interactions skill and this factor had highest percentage of explained variance (39.49%).

The second factor was largely related to teacher's teaching competency. This factor consists of 11 items (30, 33, 34, 32, 31, 21, 35, 29, 39, 20, 10). The proportion of variance explained by this factor was 5.32%. This factor taps capability of teaching to better provide teaching using good teaching strategies

The third factor consisted of 11 items (44, 45, 46, 41, 40, 47, 38, 47, 43, 32, 36). This factor was related to the assessment competency which taps the degree to which the teachers assessed students' learning appropriately. The proportion of variance explained by this factor was 4.05%.

The fourth factor consisted of 6 items (14, 13, 12, 11, 15, 8). This factor was closely related to teacher's knowledge and support tapping teachers' subject-matter knowledge and their supports to promote students' knowledge. The proportion of variance explained by this factor was 3.60%.

The fifth factor consisted of 6 items (7, 2, 3, 6, 1, 4). This factor was related to teacher's charisma that describes the general good personality and characteristics of teacher. The proportion of variance explained by this factor was 2.83%.

The last factor consisted of 2 items (5, 22). This factor was related to teacher's temperament that taps the sense of humor of teacher. The proportion of variance explained by this factor was 2.41%.

Table 4. Factor Loadings

item	Dimension					
	1	2	3	4	5	6
18 Teacher respectfully accepts students' opinions	.684					
17 Teacher motivates and empowers students	.666			.326		
27 Teacher perceives students' needs	.649					
16 Teacher has good attitude to students	.632					
24 Teacher acts as a good adviser	.619					.310
26 Teacher pays attention to students	.606	.361				
25 Teacher is available for getting helps/consults	.599		.331			
28 Teacher perceives students' differences	.593		.344			
23 Teacher is friendly	.562					.471
19 Teacher is willing to answer questions	.562	.441				
9 Teacher is kind	.472				.384	
30 Teacher covers all the contents in the syllabus		.679				
33 Teacher has well-prepared lessons		.677				
34 Teacher summarizes the main concepts		.646				
32 Teacher shows good examples to students		.601	.328			
31 Teacher selects appropriate teaching methods		.566				.309
21 Teacher is enthusiastic about teaching	.349	.557			.308	
35 Teacher explains the lessons clearly	.311	.566				
29 Teacher informs learning objectives		.543	.305			
39 Teacher has logical sequences of teaching		.535	.470			
20 Teacher devotes his/herself to teach students	.424	.504				
10 Teacher is knowledgeable		.471		.460		
44 Teacher assesses students' prior knowledge			.614	.349		
45 Teacher questions to evaluate students' knowledge			.607			
46 Teacher teaches topics students misunderstood	.312		.601	.310		
41 Teacher uses several assessment methods			.598			
40 Teacher informs how to grade students		.368	.540			
47 Teacher uses student assessment to improve teaching	.340		.535	.384		
38 Teacher uses materials appropriate to contents		.405	.506			
37 Teacher used many teaching materials			.479			

Table 4. Factor Loadings (continued)

item	Dimension					
	1	**2**	**3**	**4**	**5**	**6**
43 Teacher reports students' weakness/strength to student	.372		.467			
42 Teacher grades students fairly	.330	.372	.415			
36 Teacher encourages students to take participation		.378	.393			
14 Teacher has other class-related knowledge				.653		
13 Teacher promotes students' creativity				.653		
12 Teacher has a wide span of knowledge				.633		
11 Teacher teaches topics relevant topics being taught		.356		.625		
15 Teacher points out the class-real life relationship				.613		
8 Teacher is punctual		.346		.379		
7 Teacher is confident in teaching					.734	
2 Teacher is patient	.325				.699	
3 Teacher is responsible					.647	
6 Teacher has good relationship to students					.640	.514
1 Teacher is a reasonable person	.382				.611	
4 Teacher speaks clearly				.328	.598	
5 Teacher has a sense of humor						.668
22 Teacher creates enjoyable learning environment	.379			.303		.555

Table 5 shows the results of analysis of variance that were performed in order to compare the average of the 6 extracted dimensions evaluated by teachers, peers, and students. These dimensions would be then used later in the section of MTMM analysis. From the results of analysis of variance, ratings evaluated by teachers and peers were not statistically significant, except for the fifth and the sixth dimension. This indicated that the results of evaluation evaluated by teachers and their pees were quite similar. However, students evaluated their teacher's performance in ways that was significantly differently from teachers and peers did. Specifically, the means performance dimensions from student evaluation were statistically lower than those obtained from teachers and peers.

Figure 2 graphically displays the comparison of the performance dimensions obtained from teachers, their peers, and students. This picture supported the interpretation of the analysis of variance in table 5.

Table 5. The comparison of mean ratings across raters

Dimension	(I) Rater	(J) Rater	Mean Difference (I-J)	Std. Error	Sig.
Interpersonal Skill	student	peer	-0.63	0.06	0.00
	student	teacher	-0.72	0.08	0.00
	peer	teacher	-0.09	0.09	0.61
Teaching Competency	student	peer	-0.38	0.05	0.00
	student	teacher	-0.40	0.07	0.00
	peer	teacher	-0.02	0.08	0.96
Assessment	student	peer	-0.50	0.05	0.00
	student	teacher	-0.62	0.07	0.00
	peer	teacher	-0.12	0.09	0.39
Knowledge	student	peer	-0.46	0.06	0.00
	student	teacher	-0.51	0.07	0.00
	peer	teacher	-0.06	0.09	0.81
charisma	student	peer	-0.34	0.05	0.00
	student	teacher	1.72	0.06	0.00
	peer	teacher	2.06	0.07	0.00
Temperament	student	peer	-0.45	0.08	0.00
	student	teacher	-0.69	0.10	0.00
	peer	teacher	-0.24	0.12	0.14

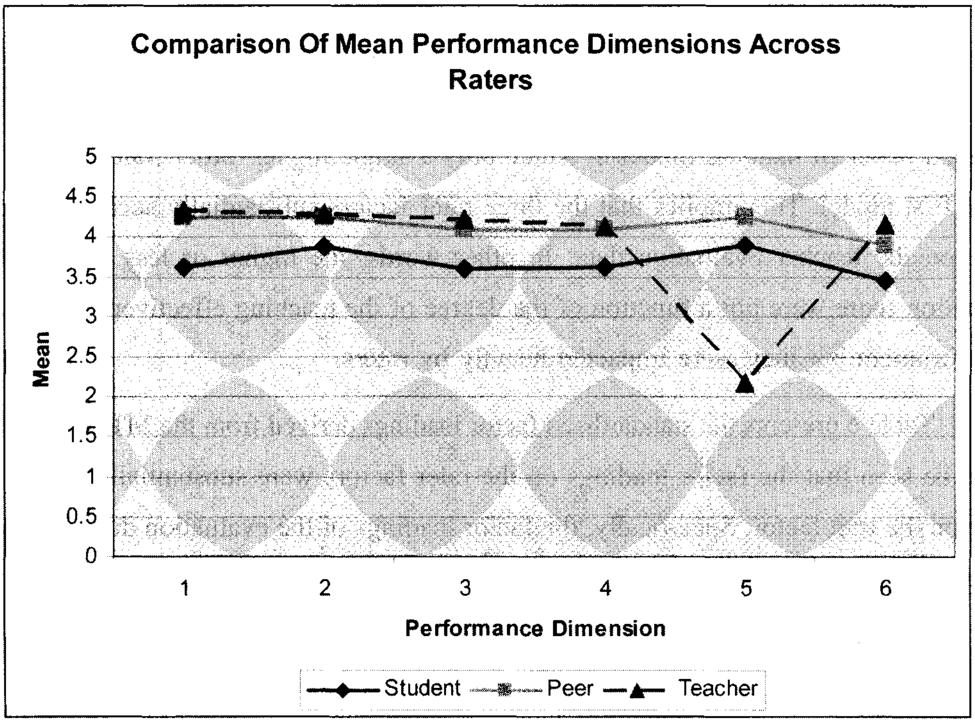

Figure 2. The comparison of mean performance dimensions across raters

Halo Effect Analysis

The analysis of the halo effect was actually the analysis of a correlation matrix in table 6. This matrix presented the correlations among the evaluation dimensions which were derived from the EFA analysis as mentioned earlier. Two models were hypothesized and tested. The first model consisted of the six evaluation dimensions and only a single higher-order trait which was the teaching effectiveness. This model implies that only a single higher-order trait factor accounted for the variation in the evaluation data. In other words, rater's characteristics did not distort the actual performance of teachers. The second model was similar to the first model but did include the three rater factors: teacher, student, and peer. This model implies that, in addition to a single high-order trait, the three rater factors also accounted for the variation in the evaluation data. These two models were analyzed separately. Then two models were compared in terms of their fit statistics. If the second model provided better fit indices and there were loadings relatively heavily loaded on the raters factors, it would indicate that halo effect existed because rater factors explained the variation in the evaluation data significantly.

The results of the analysis of MTMM were shown in table 7 and 8. Table 7 presents the chi-square and goodness-of-fit indices obtained from the analysis of the two models. It can be seen that the second model in which rater factors were included provided better fit than the model in which only a single higher-order trait was modeled in the first model. This implies that the three rater factors substantially accounted for the variation the teacher evaluation data. In other words, the higher or less values of the evaluation score were not a function of the degree of the teaching effectiveness measured the instrument but they were impacted heavily by raters.

Table 8 presents the standardized factor loadings derived from the MTMM analysis. It can be seen that the factor loadings on the rater factors were substantially higher than those on the trait factor. Specifically, the factor loadings of the evaluation dimensions that loaded on the teaching effectiveness factor as shown in the first column ranging from -.04 to .59 were relatively smaller than those that loaded on the rater factors. This indicated that the variation in the teacher evaluation data was influenced by the rater factors. This evidence suggested the existence of halo effect in the teacher evaluation data evaluated by multiple raters.

Table 6. A Correlation Matrix of Evaluation Dimensions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Self-Interpersonal skill	1.00																	
Self-Teaching	0.71	1.00																
Self-Assessment	0.62	0.75	1.00															
Self-Knowledge	0.60	0.71	0.57	1.00														
Self-Charisma	0.61	0.62	0.51	0.64	1.00													
Self-Temperament	0.44	0.29	0.33	0.48	0.69	1.00												
Student-Interpersonal skill	0.03	0.06	0.19	0.09	0.01	0.09	1.00											
Student-Teaching	0.05	0.09	0.12	0.12	0.04	0.04	0.85	1.00										
Student-Assessment	0.07	0.09	0.13	0.08	0.00	-0.01	0.85	0.83	1.00									
Student-Knowledge	0.12	0.14	0.17	0.13	0.07	0.06	0.80	0.87	0.83	1.00								
Student-Charisma	0.06	0.12	0.17	0.15	0.06	0.06	0.90	0.90	0.78	0.81	1.00							
Student-Temperament	0.06	0.07	0.14	0.10	0.07	0.13	0.80	0.67	0.73	0.72	0.81	1.00						
Peer-Interpersonal skill	0.04	0.05	0.06	-0.03	-0.04	-0.13	0.26	0.22	0.22	0.22	0.28	0.19	1.00					
Peer-Teaching	0.10	0.10	0.12	0.03	-0.14	-0.08	0.27	0.32	0.29	0.34	0.34	0.20	0.78	1.00				
Peer-Assessment	0.13	0.13	0.16	0.04	-0.02	-0.04	0.29	0.32	0.38	0.41	0.35	0.27	0.70	0.82	1.00			
Peer-Knowledge	0.08	0.06	0.11	0.00	-0.07	-0.11	0.02	-0.01	0.07	0.10	0.03	0.07	0.56	0.67	0.60	1.00		
Peer-Charisma	-0.03	0.00	0.00	0.00	-0.13	-0.09	0.09	0.06	0.08	0.12	0.13	0.13	0.75	0.63	0.59	0.66	1.00	
Peer-Temperament	0.03	0.05	0.04	0.02	-0.04	-0.08	0.02	-0.03	0.07	0.06	0.02	0.16	0.59	0.42	0.43	0.48	0.75	1.00

Table 7. Goodness-of-fit indices

Model	χ^2	df	χ^2/df	RMSE	RMR	CFI	GFI	NNFI
Trait only	1856.12	135	13.75	.36	.051	.41	.32	.33
Trait & Rater	219.99	114	1.93	.098	.012	.94	.80	.91

Table 8. Completely Standardized Loadings

	Effectiveness	Teacher	Student	Peer
Self assessment-Teaching	0.16	0.78		
Self assessment-Interpersonal skill	0.14	0.86		
Self assessment-Charisma	0.16	0.76		
Self assessment-Knowledge	0.04	0.8		
Self assessment-Temperament	0.04	0.76		
Self assessment-Assessment	-0.04	0.53		
Student-Teaching	0.3		0.90	
Student-Interpersonal skill	0.44		0.82	
Student-Charisma	0.35		0.81	
Student-Knowledge	0.39		0.79	
Student-Temperament	0.37		0.88	
Student-Assessment	0.12		0.84	
Peer-Teaching	0.25			0.83
Peer-Interpersonal skill	0.59			0.75
Peer-Charisma	0.5			0.70
Peer-Knowledge	0.19			0.72
Peer-Temperament	-0.1			0.93
Peer-Assessment	-0.27			0.78

Discussions

Halo effect seems to exist in 360-degree performance assessment. This may affect the validity of the utilization of 360-degree evaluation in teacher evaluation because there is raters' inconsistency and the inconsistency may or may not indicate the actual performance of teachers. From MTMM analysis, the analysis reveals that teacher's (actual) performances had only little effect on ratings as perceived by three types of raters, but raters themselves had more effect on ratings of teacher performance evaluation.

It was important to note that teacher evaluation were not independent from raters because the results were mixed between actual teacher performance and evaluators' attitudes. Thus, measures of teacher effectiveness were not valid. This was because the fact that to judge the degree of teachers' teaching effectiveness seems to depend largely on evaluators. For example, students were likely to provide relatively low rating results, but teachers evaluated their performance with very high ratings. Therefore, the result of 360-degree teacher evaluation seems to be limited in some degrees for some practical uses because the evaluation data may be distorted by evaluator's attitude toward teacher.

To improve the precision of data obtained through multiple evaluators, we believe that negative effects of raters might be reduced by providing training to evaluators. This should be applied to inform them prior to evaluation process about the appropriate role of an evaluator and about the aspect of performance that should be monitored and assessed.

The implication for practical uses of 360-degree teacher evaluation was consistent to the recommendation made by Mount et al (1998) in a sense that 360-degree performance assessment is appropriate for continuous improvement but not for promotion, hiring, or other administration decisions. We believe that even though halo effect is more likely to exist in 360-degree teacher evaluation, incorporating multiple evaluators in teacher assessment is worth employing because teaching effectiveness has a multidimensional nature. Different evaluators might have different views which will be useful for providing information used for improving teacher professional development. Mount and et al also suggest that the use of multiple data sources to measure teaching effectiveness can result in a fuller and more accurate views than is available through more narrowly defined approach to data collection. Therefore, the uses of teacher evaluation obtained from multiple evaluators is valuable but should be used with cautiousness.

We also suggest teachers and school principals to use data from multiple evaluators for the purpose of teacher self-assessment. In this application, it is more similar to formative assessment in that teacher may use feedback from peers and students to improve their teaching and learning. Feedback from multiple evaluators is valuable because it will provide fuller information about teacher's practices.

The limitations of this study were that there was no much training provided to evaluators in this study. Researchers simply introduced the objective of this research to them before they were asked to evaluate teachers' teaching effectiveness. To provide enough training to raters might enable researchers to obtain more firmly understanding and valid interpretation of the utilization of multiple evaluators in teacher performance evaluation contexts. In addition, this study used only one evaluation instrument which was the rating scale. Future research should include more various evaluation instruments such as portfolio and classroom observation to be used to evaluate teachers' teaching effectiveness.

References

Airasian, P. W., & Gullickson, A. (1997). Teacher-self evaluation. In J. H. Stronge *(Ed). Evaluating teaching: a guide to current thinking and best practice.* Thousand Oaks: Corwin Press.

Birenbaum, M. (2003). New insight into learning and teaching and their implications for assessment. In Mien Segers, Filip Dochy, & Eduardo Cascallar (Eds.), *Optimising new models of assessment in search of qualities and standards.* Dordrencht, The Netherlands: Kluwer academic publishers.

Chen, W., Burry-Stock, J. A., & Rovegno, I. (2000). Self-evaluation of expertise in teaching elementary physical education from constructivist perspectives. *Journal of personnel evaluation in education,* 14(1), 25-46.

Davis, R. D., Ellett, D. C., & Annunziata, J. (2002). Teacher evaluation, leadership, *and learning organizations. Journal of personnel evaluation in education,* 16(4), 287-301.

Elmier, H., Jenkins, R., & Clawford. G. (1991). The predictive validity of student *evaluation in the identification of meritorious teachers. Journal of personnel evaluation in education,* 4, 341-347.

Hu, L. T., & Bentler, P. M. (1993). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling,* 6, 1-55.

Jackson, D. I., & et al. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and psychological measurement,* 59(4), 580-596.

Johnson, B. (1997). An organizational analysis of multiple perspectives of effective teaching: implications for teacher evaluation. *Journal of personnel evaluation in education,* 11, 69-87.

Juntavech, J. (1999). *A study of agreement between self rating and other ratings of teaching effectiveness of secondary school teachers.* Published thesis, Chulalongkorn University.

Kaplan, D., & Elliott, P. R. (1997). A model based approach to validating educational indicators using multilevel structural equation modeling. *Journal of educational and behavioral statistics,* 22(3), 323-347.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guildford.

Millman, J. (1981). *Handbook of teacher evaluation.* Beverly Hills: Sage Publications.

Mount, M. K., & et al. (1998). Trait, rater and level effects in 360-degree performance *ratings. Personnel psychology,* 51(3), 557-576.

Pillay, H. (2002). *Teacher development for quality learning: The Thailand Education Reform Project.* Office of the National Education Commission, Office of the Prime Minister, Thailand.

Shepard, G. (2005). How *to make performance evaluations really work.* Hoboken, NJ: John Willey & Sons.

Stufflebeam, D.L., & Shinkfield, A. J. (1985). *Systematic evaluation.* Boston: Kluwer-Nijhoff.

Wang. M. C., Haertel, G. D., & Walberg, H. J. (1993). Towards a knowledge base for school learning. *Review of educational research,* 63(3), 249-294.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teachers and classroom context effects on student's achievement: implications for teacher evaluation. Journal of *personnel evaluation in education,* 11, 57-67.