Sampling Strategy in Evaluating Teaching Performance through Student Ratings[1]

James Roldan S. Reyes
Institute of Statistics, University of the Philippines Los Baños

Zita VJ. Albacea
Institute of Statistics, University of the Philippines Los Baños

This paper presents an alternative method apart from the current online or electronic approach, which is currently being used by some higher education institutions (HEIs), in administering student ratings for teachers. The developed method still employed the traditional paper approach but has been improved through the use of sampling application which includes sampling design, sample size, estimation technique, and strategic implementation. Three basic sampling designs such as simple random, stratified random, and cluster sampling were applied at three different sampling rates such as 25%, 50%, and 75%. For the empirical evaluation of the developed method, the Student Evaluation of Teachers (SET) of the University of the Philippines Los Baños (UPLB) was utilized using bootstrap resampling technique. Based on findings, stratified random sampling is the most appropriate sampling design to use with 50% of the students for each class section serving as SET evaluators. Results also revealed that bootstrap estimates of standard error are lower than that of the standard error using jackknife resampling procedure. Generally, the improved traditional paper approach same with the electronic approach could reduce the cost of administering student ratings. However, the electronic approach has a dilemma with regards to high non-response bias leading to invalid results. Thus, to minimize non-response error of the developed method, its standard protocol to administer the student ratings has been formulated.

Keywords: student ratings, traditional paper approach, sampling application, bootstrap resampling, ackknife resampling, nonresponse error

1 Major results presented in this paper were taken from the unpublished Master's thesis of JRS Reyes under the supervision of ZVJ Albacea.

1. Introduction

Organizational output's quality is greatly impinged on any input contributed by its members. Poor work performance of its members per se could directly withhold the purpose of the organization towards its success. In relation to this, each organization, which includes academic institutions, systematically conducts evaluation of employee's work performance to determine who among their staff should be retained, promoted, or terminated. A good measure in assessing teaching performance aid academic institutions to acquire quality output, the students, through quality input, the teachers. Some of these measures include student ratings, peer ratings, self-evaluation, learning outcomes, and teaching portfolio.

Among these measures, the rating given by the students or student rating is the most prevailing measure since higher education institutions (HEIs) use this kind of measure to come up with decisions such as on teachers' promotion and tenure. This performance measure also lies on the principle that the students are the direct customers of the services offered by the teachers and hence, they are deemed to be the best evaluators to assess their teachers' performance.

Many studies have been published which supported the validity, reliability, and usefulness of student ratings. In fact, there is more research on student ratings than other topics in higher education (Theall & Franklin, 1990). McKeachie (1997) defined student ratings as the single most valid source of data on teaching effectiveness. Furthermore, Aleamoni (1987) mentioned it as a form of communication encouragement between student and teacher that could further lead to their involvement in the teaching-learning process which could later lift the level of instruction. Oreovicz and Wankat (1993) defined and assessed the reliability and validity of this teaching performance's measure. Reliability refers to the stability of ratings while validity means ratings can measure what it supposed to measure. Using correlation analysis, the study found quite high internal consistency or agreement of student ratings in the same class. Moreover, it is revealed that student ratings have modest positive ratings when compared to the other methods of evaluating teaching performance.

A proper administration of student ratings is considered to be of great importance since it could affect student assessment of teachers (Marsh, 1984). According to Coburn (1984), informal administration of student ratings could lead to bias in ratings, low return of the forms, or less candid responses on the part of the students. Student ratings' bias related factors include instructor's presence in the room, time of administration, student anonymity, and instructions in the evaluation form (Jacobs, 2002). Students might give a high rating with the faculty present during the evaluation. A positive bias might also be present for the faculty evaluated earlier in the semester than those evaluated near the scheduled release of final grades. In the study about teaching practices in colleges and schools of pharmacy by Barnett and Matthews (2009), it is found that majority

of colleges or schools administered the student ratings at near the conclusion of the course while only about 38% administered it at the middle of the semester. The advance announcement of the evaluation schedule to the students could also influence the result since only those students who want to evaluate their teachers might be present during the time of evaluation. However, Siegfried and Vahaly (1975) presented evidence that announcing the schedule of evaluation does not introduce any particular bias on the results of student ratings. In fact, Davis (1988) suggested a procedure in administering student ratings which includes the advance announcement of the date in which rating forms will be handed to the students. Other suggestions includes the following: (1) informing the students about the purpose of the evaluation; (2) asking students to complete the form anonymously; (3) designating the person-in-charge of administration; and (4) looking at the results only after all grades for the course are submitted. The suggestions made by Davis are the procedures a teacher needs to follow in administering student ratings. However, teachers should not be involved in administering the student ratings since any teacher's verbal comment whether it is slight or obvious to the students could introduce bias (Marsh, 1984). A teacher who makes slight comments pleading for a high rating for his or her own good might be able to receive favorable student ratings. To avoid these bias related factors, a standard procedure of administering student ratings should be followed. However, different HEIs have still varied systems in handling their student ratings. Among these systems used by HEIs, the absence of the evaluator and the confidentiality of the evaluator's identity are the two most common provisions. In fact, research shows that requiring students to sign the forms inflates the student ratings (Cashin, 1990) and added student's concerns about the possibility of retribution (Ory, 1990).

On the other hand, for any appraisal method, cost is an aspect at all times. Cost includes the reproduction of materials, payment for the staff who are involved in the administration and in processing of the results, and the machines to be used for data processing. Higher number of evaluators in the performance evaluation means a larger cost. Burnett and Verma (1996) conducted a study on how to cut down the cost of evaluation systems. The study proposed the reduction of number of evaluators by reducing the sample size without undermining the validity of the evaluation results. Findings showed that the cost of data gathering and entry could substantially reduce from $1,116 at 1% error to $348 at 3% error.

Currently, most HEIs involve all students who are present during the time of evaluation as evaluators of student ratings. For a large university with a large population of students and teachers, this is not fairly cost-effective as well as not managerially effective. A lot of forms and vast amount of data to process require cognitive weight to human that makes student ratings error prone. Hence, it is very important to reduce the manpower involvement in administering the student ratings for teachers (Burnett and Verma, 1996).

Some universities are currently administering their student ratings via online or electronic approach. In a comparative analysis between paper and online administration of student ratings, evaluation scoring patterns were similar for both methods, however, student response for online approach was lower (Fike, et al., 2010). Although online administration could definitely decrease a large amount of cost, the presence of a high non-response rate which could lead to a non-representative sample is a major dilemma. Moreover, these non-representative samples could lead to invalid conclusion of the results. Hence, a representative sample through sampling application is thought of. Instead of involving all the students in the evaluation process or employing electronic approach, the selection of a random sample of students as evaluators is considered.

In University of the Philippines Diliman (UPD), a 50% sample of students in each class of the teacher to be evaluated is randomly selected by the University Research Assistant (URA) given that the class size is above fifteen students. The sampling process is done by choosing a random start from the student and asking the other students to count off 1, 2, 1, 2 … following a path around the room. In deciding which group of students will be the evaluators of the student ratings, the URA tosses a coin. If the coin turns up head, group 1 will be asked to stay and answer the student ratings while group 2 will be advised to leave the room and vice versa (http://www.ovcaa.upd.edu.ph). This method of administering the student ratings is efficient in terms of cost reduction and easier data processing.

Despite the recent application of sampling in the evaluation of work performance, particularly the teaching performance, generation of sample students to be part of the student ratings is still not being employed by most HEIs. An alternative method of administering the students using random sample of students, as student ratings' evaluators, must be looked into. This paper presents results of a study which aimed to improve the traditional paper approach with the use of sampling application. The cost-efficiency of the proposed sampling methodology is being compared from the traditional paper approach and not with the electronic approach. It is true that electronic approach is more cost-efficient than the proposed sampling methodology. However, the effect of possible measurement error and high non-response bias using the electronic approach must be considered. In proposing an alternative approach in administering the SET, a balanced between its cost-reduction and reliability and validity of results must be taken into account.

In the University of the Philippines Los Baños (UPLB), the university regularly exhausts per semester a lot of resources in terms of money, effort, and time in administering their student ratings known as Student Evaluation of Teachers (SET). The consumption of millions of papers and computing machines undeniably incur high amount of cost for the university. With around 12,000 students per semester and more or less 39 administrative staff involved in SET administration, the usual annual amount spent in SET is approximated to be

equal to 130,000 pesos. Also, an average of 5,000 forms per day is processed in terms of scanning. Same as the electronic approach, the application of sampling in administering the SET could trim down the cost expended for it and might be used for some other university's purposes. Even the work efforts handed by administrative staff, who is the person-in-charge of the SET administration, will be lessened through the use of sampling and their efforts rechanneled to their other assignments. With the selection of sample students as evaluators, a more efficient administration of the SET and more accurate student ratings' results will be obtained since the administrative task of handling the SET is at once less demanding than that of requiring all students to evaluate their teachers. Also, the add-on of sampling to the traditional paper approach would also provide reduction of human involvement providing a lesser risk of having non-sampling errors, such as measurement error and processing error, which could later lead to biased results. Furthermore, a more speedy release of results and reports will be realized with the application of sampling on the student ratings.

A scientific-based decision is based on accurate information and in order to generate an accurate information on student rating, the following are considered: (1) the appropriateness and accuracy of the sample collection and handling method; (2) the effect of possible measurement error; (3) the quality and appropriateness of the statistical analysis; and (4) the representativeness of the sampled SET scores with respect to the objective of the study. In the development of every sampling strategy that adequately addresses the estimation or decision at hand, it is necessary to understand what relevant factors should be considered and how these factors affect the choice of an appropriate sampling strategy.

In UPLB, there is only one standard form of Student Evaluation of Teachers (SET) that is being used despite of the course and college from which a particular faculty is being evaluated. Furthermore, the defined population of the study is the set of all students handled by the faculty for a particular semester from which a faculty from UPLB, the domain of estimation, belongs only to one college. In this regards, courses and colleges as grouping variables for the some basic sampling designs are not considered. In the development of the sampling methodology for UPLB SET, the homogeneity of individual units of the defined population is taken into account. Homogeneity refers to the uniformity of units on all characteristics that could affect the variable of interest. However, the environment to which the students are subjected to make the defined population heterogeneous. In this regard, students might be grouped according to class section since the characteristics of students in terms of evaluating the faculty might be the same within section but varying among sections making the students within class section homogenous. In fact, teaching performance has found to be significantly different across sections of the same course taught by a faculty (Persons, 2007). Also, the presence of auxiliary variables in the frame such as section, subject, and number of students alleviate the development of the proposed sampling methodology.

Moreover, the appropriateness of sampling process also depends on the size of the population since too small population impedes the sampling application. The geographical distribution of the units must be considered in the development of sampling plans. In controlling measurement error, the widely distributed sampling unit implies a complex design. Thus, since students to be sampled are easily found inside the classroom, the appropriateness of a more basic design is realized. Also, for the ease of administering the SET without requiring for a technical person to be present in administration, basic sampling designs are considered. These are the basis for considering three of the basic sampling designs namely simple random sampling, stratified random sampling, and cluster sampling.

2. Methodology

2.1 Data

For the availability of data, the results of SET of the UPLB faculty during the 1st Semester 2009 - 2010 are used as test data. Also, the sampling frame used is the list of all UPLB faculty for the same semester who participated in the SET.

2.2 Sampling design, sample size, and estimation technique

The auxiliary variables such as class size and class section, which were available in the frame, were used to divide the SET scores according to class mode with five groups such as large lecture class mode (with a minimum class size of 120 students), lecture class mode, laboratory class mode, recitation class mode, and small class mode (with class size below 10 students). For each class mode excluding the small class size wherein a census of SET scores is considered, three basic sampling methods such as simple random (SRS), stratified random (StRS), and cluster sampling are considered. For StRS and cluster sampling, the data on SET scores is then grouped according to class section. For the sample size, different sampling fractions at 25%, 50%, and 75% rates are generated from each sampling method. Sampling fraction is used instead of via precision point of view or formula approach for the ease of administering the SET without requiring for a technical person to be present in actual administration. Such could also add cost. The estimates of the true weighted average SET scores with their corresponding standard error are computed for each of the sampling methods considered.

2.3 Empirical evaluation

For the evaluation of the different properties of the considered estimator and for the determination of the optimum sample size, bootstrap resampling technique is applied. From the population SET data, sampling with replacement is done considering the combinations between the three basic sampling designs and three different sampling fractions. The sampling is done in several replications but due to limitation of computing resources, only a thousand replications, b,

were generated instead of all possible resamples. The bootstrap estimate $\hat{\theta}_b$, is calculated for each replication based on the design used. The probability distribution from b $\hat{\theta}_b$'s by placing a probability of $1/b$ for each $\hat{\theta}_b$ is constructed. This distribution is the bootstrap estimate of the sampling distribution of $\hat{\theta}$ for a particular design. This sampling distribution of the estimator using the bootstrap estimates was assessed and their mean and standard error were computed for each combination of sampling design and sampling fraction. The sampling distribution of the estimator, its accuracy in terms of bias, and its precision based on standard error are assessed and compared with the true weighted average SET scores. In choosing the appropriate sample size, the sample size for which the largest percent gain in mean square error (MSE) is observed as the appropriate sample size. The percentage gain in MSE is computed as:

$$Gain(\%) = \left| \frac{MSE_{t-1} - MSE_t}{MSE_{t-1}} \right| \times 100$$

where $t-1$ is the previous sampling rate and t is the current sampling rate. Lastly, the bootstrap estimates of the standard error are then compared with the standard error using another resampling technique that is Jackknife. All computations were done using licensed STATA (version 12).

3. Results and Discussions

3.1 The Student Evaluation of Teacher (SET) population data

A total of 71,573 SET scores are considered in the study. Descriptive statistics of SET scores found the minimum and maximum SET scores of 1.000 and 3.000, respectively, wherein these two scores are the ceiling scores that a faculty might obtain using the UPLB SET. A faculty with score closer to 1.000 has a higher SET score compared to a faculty with score closer to 3.000. The SET scores tend to cluster at μ_N =1.497. It also exhibits large variability relative to the mean with a coefficient of variation equal to 19.756%. In terms of symmetry, students usually give high SET scores than low SET scores. The SET score was then correlated with class size and class mode. It was found that there is a direct very weak relationship between the SET score and class size with a Spearman correlation coefficient of 0.110. Since in UPLB SET, a faculty with score closer to 1.00 has a higher SET score compared to a faculty with score closer to 3.00, the correlation coefficient revealed that as the class size increases the SET score of a faculty decreases. On the other hand, in correlating the SET score with class mode, the SET scores were categorized into two such as Good and Poor. Scores that fall in the Good category are SET scores between 1 and 1.75; otherwise Poor Category. Using chi-square-based measure of association, it was found that the SET score

and class mode have a moderate relationship with Cramer's V coefficient of 0.124. With these results, the SET scores are grouped into five according to class mode such as large lecture class, lecture class, laboratory class, recitation class, and small class.

3.2 Developed sampling design, sample size, estimation technique, and strategic implementation

Tables 1 to 4 show the comparison in terms of statistical properties of the estimator in using different sampling designs and different sampling rates for each class mode. Results showed that the bias decreases as the sampling rate increases for classes with small number of students such as laboratory and recitation class modes. This implies that the estimator tends to be more accurate as the number of sample students increases. However, there is no perceived pattern between the bias and the sampling rate for lecture and large lecture class modes. The study also found that there is a direct relationship between the standard error and the sampling rate across class mode. This means the estimator tends to be more precise as the sampling rate increases. Results also showed that the estimate of SRS usually produces the most accurate estimate of the parameter, μ_N, for all class modes as measured by the bias. This means that in the long, the expected value of the means of all possible simple random samples is much closer with the parameter as compared with other two sampling designs. However, in terms of precision, the estimate of StRS consistently gives the most precise estimate across class mode. This means that the means of all possible stratified random samples are closer to each other compared to the means of all possible simple random and cluster samples. With the estimator of SRS and StRS as the most accurate and precise estimators, the mean square error (MSE) is computed. This measure depicts the overall measure of accuracy and precision. Results found that across sampling rate and across sampling design, StRS is found to be the most appropriate sampling method for the UPLB SET since it gives lower MSE. The lower the MSE, the more accurate and precise the estimator is.

For the determination of the optimum sample size, the percentage gain in MSE is then computed. Generally, a half decrease in MSE is obtained when the sampling rate is changed from 25% to 50% while only a quarter decreased is gained when the sampling rate is changed from 50% to 75% as shown in Table 5. Thus, the study revealed that the selection of a half number of students in each class section already represents the whole population of students to evaluate the teaching performance.

On other hand, due to a very small number of sampling units, a census of students in a small class mode is considered.

Furthermore, the developed method was assessed for some faculty representing a particular type and it was empirically observed that the proposed estimator is numerically unbiased and precise having a small variance of the estimate per

class mode. For example, a faculty who handled all laboratory classes for the semester obtained a SET estimate score of 1.45372 with variance of 0.0010 using the developed method which is the same with its true value.

Table 1. Statistical Properties of the Estimator at Different Sampling Rates for the Three Basic Sampling Methods in Large Lecture Class Mode

Sampling Design	25% Sampling Rate	50% Sampling Rate	75% Sampling Rate
Simple Random Sampling			
Bias	-0.00002	0.00006	0.00006
Standard Error	0.00279	0.00193	0.00162
Mean Square Error	0.00001	<0.00000	<0.00000
Stratified Random Sampling			
Bias	-0.00011	-0.00016	-0.00005
Standard Error	0.00226	0.00161	0.00136
Mean Square Error	0.00001	<0.00000	<0.00000
Cluster			
Bias	-0.00015	0.00006	0.00011
Standard Error	0.01098	0.00773	0.00625
Mean Square Error	0.00012	0.00006	0.00004

Table 2. Statistical Properties of the Estimator at Different Sampling rates for the Three Basic Sampling Methods in Lecture Class Mode

Sampling Design	25% Sampling Rate	50% Sampling Rate	75% Sampling Rate
Simple Random Sampling			
Bias	-0.00021	-0.00004	-0.00003
Standard Error	0.00459	0.00322	0.00264
Mean Square Error	0.00002	0.00001	0.00001
Stratified Random Sampling			
Bias	-0.00037	0.00003	-0.00008
Standard Error	0.00341	0.00242	0.00218
Mean Square Error	0.00001	0.00001	<0.00000
Cluster			
Bias	0.00012	0.00014	0.00003
Standard Error	0.01042	0.00733	0.00590
Mean Square Error	0.00011	0.00005	0.00003

Table 3. Statistical Properties of the Estimator at Different Sampling Rates for the Three Basic Sampling Methods in Laboratory Class Mode

Sampling Design	25% Sampling Rate	50% Sampling Rate	75% Sampling Rate
Simple Random Sampling			
Bias	-0.00033	-0.00004	-0.00007
Standard Error	0.00837	0.00591	0.00484
Mean Square Error	0.00007	0.00003	0.00002
Stratified Random Sampling			
Bias	-0.00050	-0.00023	-0.00028
Standard Error	0.00784	0.00537	0.00441
Mean Square Error	0.00006	0.00003	0.00002
Cluster			
Bias	0.00011	-0.00025	0.00012
Standard Error	0.04609	0.03204	0.02605
Mean Square Error	0.00212	0.00103	0.00068

Table 4. Statistical Properties of the Estimator at Different Sampling Rates for the Three Basic Sampling Methods in Recitation Class Mode

Sampling Design	25% Sampling Rate	50% Sampling Rate	75% Sampling Rate
Simple Random Sampling			
Bias	-0.00013	-0.00006	-0.00003
Standard Error	0.00909	0.00645	0.00526
Mean Square Error	0.00008	0.00004	0.00003
Stratified Random Sampling			
Bias	-0.00066	0.00033	0.00007
Standard Error	0.00777	0.00516	0.00431
Mean Square Error	0.00006	0.00003	0.00002
Cluster			
Bias	0.00086	0.00084	0.00075
Standard Error	0.02122	0.01482	0.01174
Mean Square Error	0.00045	0.00020	0.00014

Table 5. One Sample Point-gain in MSE (%) using the Three Sampling Fractions for the Different Class Modes

Class mode	Simple Random Sampling	Stratified Random Sampling	Cluster
Large Lecture Class			
25% vs. 50%	50.28	53.40	51.67
50% vs. 75%	32.66	32.29	33.88
Lecture Class			
25% vs. 50%	52.56	47.06	50.41
50% vs. 75%	29.73	29.63	34.62
Laboratory Class			
25% vs. 50%	50.71	50.85	50.55
50% vs. 75%	32.69	18.97	35.20
Recitation Class			
25% vs. 50%	49.52	55.99	55.19
50% vs. 75%	33.81	30.60	35.20

Table 6 shows the comparison of the bootstrap estimates of the standard error with the standard error using another resampling technique which is known as Jackknife. It can be seen that the standard error obtained using bootstrap resampling are usually lower than that of the standard error using jackknife resampling for the different sampling rates and sampling designs. However, for recitation and laboratory class modes using SRS, standard errors of bootstrap resampling are consistently higher than that of jackknife.

A standard protocol in administering the SET for the improved traditional paper approach with sampling application is then formulated. A 50% predetermined list of simple random sample of students for each class section will be generated. This predetermined list is strictly confidential wherein nobody is allowed to see the generated list except for the administrative staff who is assigned for the SET administration. To minimize non-response error, a ten percent additional sample for replacement will be also generated. This sample replacement must be used only if the original sample is not present during the time of evaluation. The order of the list of sample replacement must be followed wherein the first replacement to be used will be the first student in the list. To minimize other non-sampling error such as measurement error that could also contribute bias, the teacher's presence in the room is strictly prohibited.

SET must be administered for about 15 to 20 minutes before the start of the class to avoid the effect of just recalling the latest teaching performance of the faculty being evaluated. Also, it must be administered during less than a month before the end of the class or a week before the end of the lecture for team teaching subjects. For recitation and laboratory class modes, the names of the sampled students will be announced inside the room and the sampled students

Table 6. Comparison Between the Estimates of the Standard Error Using Bootstrap and Jackknife Resampling Techniques

Sampling Design	25% Sampling Rate	50% Sampling Rate	75% Sampling Rate
Large Lecture Class Mode			
Simple Random Sampling			
Bootstrap s.e.	0.00837	0.00591	0.00484
Jackknife s.e.	0.00865	0.00606	0.00494
Stratified Random Sampling			
Bootstrap s.e.	0.00784	0.00537	0.00441
Jackknife s.e.	0.00849	0.00649	0.00584
Cluster			
Bootstrap s.e.	0.04609	0.03204	0.02605
Jackknife s.e.	0.08254	0.04266	0.03659
Lecture Class Mode			
Simple Random Sampling			
Bootstrap s.e.	0.00279	0.00193	0.00162
Jackknife s.e.	0.00291	0.00203	0.00165
Stratified Random Sampling			
Bootstrap s.e.	0.00226	0.00161	0.00136
Jackknife s.e.	0.00236	0.00193	0.00171
Cluster			
Bootstrap s.e.	0.01098	0.00773	0.00625
Jackknife s.e.	0.01280	0.00954	0.00824
Laboratory Class Mode			
Simple Random Sampling			
Bootstrap s.e.	0.00459	0.00322	0.00264
Jackknife s.e.	0.00428	0.00308	0.00252
Stratified Random Sampling			
Bootstrap s.e.	0.00341	0.00242	0.00218
Jackknife s.e.	0.00365	0.00278	0.00246
Cluster			
Bootstrap s.e.	0.01042	0.00733	0.00590
Jackknife s.e.	0.01168	0.00915	0.00826
Recitation Class Mode			
Simple Random Sampling			
Bootstrap s.e.	0.00909	0.00645	0.00526
Jackknife s.e.	0.00890	0.00634	0.00522
Stratified Random Sampling			
Bootstrap s.e.	0.00777	0.00516	0.00431
Jackknife s.e.	0.00797	0.00596	0.00527
Cluster			
Bootstrap s.e.	0.02122	0.01482	0.01174
Jackknife s.e.	0.02993	0.02353	0.01940

in a class section will be asked to answer the evaluation form. Students who are not part of the evaluation are required to stay inside the room. However, for large lecture and lecture class modes, the names of the sampled students will be flashed on the screen and the same procedure with the recitation class and laboratory class modes applies. Table 7 shows the strategic plan for the implementation of the developed method which includes the activities involved in the implementation and recommended schedule.

Table 7. List of Activities with Their Corresponding Schedule for the Implementation of the Developed Method

Activity	Schedule
Generation of predetermined sample	A month before the middle of the semester
SET data collection	A month(week) before the end of the semester (before the end of the lecture for team teaching subjects) teaching subjects
Processing of SET results	End of the semester

4. Conclusions

As a result, this study established the efficacy of sampling in evaluating work performance of the staff in particular to teaching performance through student ratings. Based on findings, the sampling application for the traditional paper approach involves stratified random sampling with class section as stratification variable at 50% sampling rate. For UPLB, this developed method of administering student ratings could reduce cost into half from the annual 130,000 pesos to approximately 60,000 pesos per year. For its strategic implementation, a predetermined list of simple random of students for each section will be selected. Also, in order to minimize non-sampling errors, its standard protocol is then formed.

5. Limitations and Future Directions

This study is delimited to improve the traditional paper approach in administering student ratings with reduced cost and without compensating for the results. This paper does not intend to compare the cost-efficiency between the electronic approach and the developed method. Also, coming up with current estimates of the SET scores and evaluating the reliability and validity of the existing instrument used in UPLB are not included.

Moreover, further study is also suggested to evaluate the respondents' fatigue as sample evaluators and determine its effect on the student ratings' results. Moreover, the methodology employed could also be used not only in the development of new method in evaluating work performance in the academe but also in evaluating work performance in other industry.

REFERENCES

ALEAMONI, L.M., 1987, *Handbook of Teacher Evaluation: Student Rating of Instruction*, Beverly Hills, CA, 110-145.

BARNETT, C.W. and MATTHEWS, H. W., 2009, Teaching evaluation practices in colleges and schools of pharmacy, *Am J Pham Educ* 73(6).

BURNETT, M.F. and VERMA S., 1996, Cutting evaluation costs by reducing sample size, *Extension Journal, Inc* 34(1).

CASHIN, W. E., 1990, *Student Ratings of Teaching: Recommendations for Use*, Manhattan: Center for Faculty Evaluation and Development in Higher Education, Kansas State University, 22.

COBURN, L., 1984, ERIC clearinghouse on test measurement and evaluation, Available at http://ericdigests.org/pre-927/student.htm.

DAVIS, B.G., 1988, *Sourcebook for Evaluating Teaching*, Berkeley: Office of Educational Development, University of California.

FIKE, D.S, DOYLE, D.J. and CONNELLY, R.J., 2010, Online vs. paper evaluations of faculty: When less is just as good. Available at http://uncw.edu/cte/et/articles/Vol10_2/Fike.pdf

JACOBS, L.C., 2002, Student ratings of college teaching: What research has to say, Available at http://www.indiana.edu/~best/multiop/ratings.html.

MARSH, H.W., 1984, Student's evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility, *J. Educ Psychol* 76.

MCKEACHIE, W. J., 1997, Student rating: The validity of use, *American Psychologist*, 52: 1218-1225.

OREOVICZ, F.S. and WANKAT P. C., 1992, *Teaching Engineering*, Mcgraw-Hill College, 311-314.

ORY J., 1990, Student ratings of instruction: Ethics and practice, *New Directions for Teaching and Learning* 63-74.

PERSONS, O.S., 1997, The effects of different sections and students' pre-course interest on an instructor's teaching evaluations, *Journal of College Teaching and Learning* 4: 51-56.

SIEGRIED, J. J. and VAHALY J., 1975, Sample bias of unannounced student evaluations of teaching, *Journal of Economic Education* 137-139.

THEALL, M. and FRANKLIN J. L., 1990, *Student Ratings of Instruction: Issues for Improving Practice*. San Francisco: Jossey-Bass, 43.

UNIVERSITY OF THE PHILIPPINES DILIMAN, Student Evaluation of Teachers, Available at http://ovcaa.upd.edu.ph