

Impact and treatment of the evaluators' effect on employees' performance appraisal

Awoke Seyoum Tegegne

Statistics department, College of Science, Bahir Dar University, Bahir Dar, Ethiopia

Email address:

bisrategebrail@awokeseyoum.com

To cite this article:

Awoke Seyoum Tegegne. Impact and Treatment of the Evaluators' Effect on Employees' Performance Appraisal. *Science Journal of Applied Mathematics and Statistics*. Vol. 1, No. 4, 2013, pp. 30-37. doi: 10.11648/j.sjams.20130104.11

Abstract: Putting a performance appraisal scheme is important to assess the gap between best performer and least performer employees. Employees who want to improve their work efficiency can then be rewarded, where as corrective action can be taken against those employees who don't want to improve their performance. The objective of this study is to construct a technique that helps to evaluate the subjective effect that a given evaluator's assessment will have a certain impact on the performance appraisal of a given employee, assuming that an assessment of one's work performance will have to be undertaken by an evaluator and that this assessment is essentially a subjective one. For this study, a linear mixed modeling approach will be applied to show significant evaluator's effect on a certain employees that needs to be properly accounted for when rewarding employees. With this adjustment being done, any incentive scheme, whether its motive is reward based or penalty fail in its intended purpose of improving employees' overall performance.

Keywords: Evaluators' Effect, Performance Appraisal, Model Diagnostics, Mixed Model, Fixed Effect, Best Linear Unbiased Estimator

1. Introduction

Semi-annually performance reviews are seen as critically important for ensuring the success of public entities and private companies[21]. Their aim is to induce workers to become more efficient and effective [11], and help evaluators to become more transparent in the way they interact with their employees. As a result, employees begin to have a better understanding of their evaluators' expectations, leading to a greater sense of ownership of their duties and thus improved work performance. Ignoring these performance issues will ultimately decreased morale, which in turn will lead to a drop- off the company's overall level of performance as management wastes time rectifying what isn't being done properly [6]. Thus an effective performance appraisal can provide huge benefits for the employer in terms of increased staff productivity, knowledge, loyalty and participation[13].

How one best measure the performance of an employee, however, can be significantly affected by unfavorable first impression of individuals known as evaluators? Ideally one would like to minimize the effect of this first impression on a final evaluation process, but this selective perception bias has been observed in the behavior of all evaluators, and is

therefore known as evaluators' effect [26]

Due to the complexity of the job performance and interpersonal relations at work, much of the existing research typically indicates that evaluators account for significant proportions of the variance in employees' true performance[25]; [8]. It is therefore in the interest of both the organization and the individual to maximize the effectiveness of performance appraisal by reducing the evaluator's errors. Most of the studies focus on the evaluation strategies before the evaluating rather than attending to evaluation outcomes.

Therefore, the purpose of this study is to introduce a statistical method to (i) demonstrate the possible source factors at the performance appraisal scheme; (ii) identify and adjust for the magnitude of evaluators' effect and thereby rank the 'highest' and 'least' performers, (iii) identify abnormal ratings. Hence, this study contributes to the literature by attempting to clarify the structure of evaluators' effect, the existence and nature of evaluators' effect, and the relative proportion of variance accounted for by the evaluators' influence on performance appraisal scheme.

2. The Data and Purpose of the Analysis

2.1. Study Design

The target population in which study focused is all employees and evaluators in North western part of Ethiopia. According 2009 housing and people survey of Ethiopia, the study area had around 5,500,000 employees. To do this, what I did is take random sample among government and non-government organizations and conduct cross-section study design. A total 214 employees have been selected in the study area and included in the study as each employee was part of a twice per annum based performance appraisal scheme. For each project (of activity) in which he/she was involved, that employee was given a rating on a continuum scale ranging from 0 to 25, with a higher rating showing a better performance. The evaluations were performed by 70 evaluators. The scale of complexity of the given tasks that the employees were being asked to perform was also taken in to consideration when the rating was being done by the evaluators.

2.2. Variable of Interest

The response or outcome variable for this study is the performance appraisal of employees which can be affected by rating scale of the respected evaluators. The explanatory variable for this study is evaluators effect. To alleviate the effect of different evaluators, all 70 evaluators received some form of training (i) to familiarize themselves with the measures that they would be working with, (ii) to ensure that they understood the sequence of steps that they would have to follow in their assessment and (iii) to explain how they should interrupt any normative data that they would be given. If one were able to use all 70 evaluators to rate each

and every employee in the firm, evaluators' training would minimize the effects, as the effects would be the same [17], [10]. No single employee would run the risk of having a lower or higher overall rating as all the employees would receive the same benefit or penalty from the evaluator's subjective leniency or harshness. In the firm that the study has conducted, however, not every employee was able to be rated or evaluated by the same set of evaluators. In particular, Table 1 shows how some evaluators evaluated several employees where as others only evaluated a few employees. It should be noted that in Table 1 there are 308 evaluation assessments of 214 employees because some employees were involved in a number of projects (or activities) and accordingly had multiple evaluators.

The difference between the rating that will be assigned by a single evaluator and the average rating that will be assigned by all 70 evaluators is called the 'evaluators' effect'. Clearly, if this evaluators' effect is non zero, then employees that have been evaluated by a different set of multiple evaluators may receive an unfair (i.e. biased) score first and foremost because they have faced a relatively lenient or relatively harsh set of judges when compared with the other employees in the firm. In this case, an adjustment to a given employee's average score should be made, which takes in to account the potential bias that may arise because a different set of evaluators has been used. Simply averaging the score given by each evaluator to an employee will not adjust this evaluators' effect. In the next section, it will be developed a method that attempts to account for evaluators' effect. Once this has been done, we can then separate 'highest' performers from 'least' performers and reward them accordingly.

Table 1. The number of employees per evaluator.

The number of employees per evaluator	1	2	3	4	5	6	7	8	9	10	11	13	15	16
The number of evaluator	20	10	8	5	5	3	6	3	1	4	2	1	1	1

3. Formulation of Statistical Model

A classical example of testing for inter-rater reliability is described [5] in the context of a medical situation where depressive patients are being evaluated by several psychiatrists, and there is a restriction on the number of examinations that a patient can undergo. However, this method cannot be used in this context of performance appraisal scheme because the evaluator who is evaluating a given employee is someone who has detailed knowledge of that employee's performance, i.e. the random assignment of employees to any given evaluator is not possible in this context. Furthermore, one is not necessarily able to restrict the number of employees that each evaluator sees, or vice versa.

Some researchers have suggested that one calculate a mean performance score for each employee and then rank

the employees based on their mean performance. As has already been noted, because the set of evaluators being used differs from one employee to the next, simply ranking the mean performance scores of each employee will not remove the evaluators bias in this procedure [19] other researchers have attempted to develop analysis of variance-based on raw scores [2], [10] & [26]. Such a model however required that one make use of a Likert scale when rating an employee's performance (like Excellent, very good, Good, fair, poor). In this modeling context the evaluating scheme that is given is not based on a Likert scale. In order to develop a performance score for a given employee and to correct this score for a possible evaluator's effect, I used a linear mixed model defined as shown below.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

Where Y_{ij} denotes the appraisal score of the i^{th}

employee that has been evaluated by evaluator j , μ denotes an overall mean score, α_i denotes a variation of performance of employee i from this overall mean score, β_j denotes the j^{th} evaluator's effect and ε_{ij} is an error term. In particular, it will be assumed that the α_i 's $\sim N(0, \sigma_1^2)$ i.e. α_i 's are independently, normally and identically distributed normal random variables with mean 0 and variance σ_1^2 and the $\varepsilon_{ij} \sim N(0, \sigma_0^2)$ which are also normally, independently and identically distributed normal random error terms with mean 0 and variance σ_0^2 , respectively. Focusing on the model parameter β_j , some of the management group may want to look only at the 70 evaluators, in which case the evaluators' effect β_j should be treated as being a fixed effect. On the other hand, some may argue that the 70 evaluators are representatives from a population of evaluators in the study area, in which case the evaluators' effect should be treated as being a random effect.

Instead of arguing about whether this evaluators' effect should be fixed or random, it will be constructed two statistical models: one with a evaluators' effect that is fixed where we treat this evaluator's effect β_j as $\beta_j \sim N(0, \sigma_2^2)$ being an independently, normally and identically distributed normal random variable with a mean 0 and variances σ_2^2 , it will be also assumed that α_i , β_j and ε_{ij} are distributed normally & independently of each other. The resulting model then becomes a linear random effects model. A detailed discussion about linear random effect models can be found in different researches such as [7], [18] & [20]. The main focus of interest in this model is the variance of the evaluators' effect, σ_2^2 . If $\sigma_2^2 = 0$, then the data supports the hypothesis that the evaluators' effect is constant or identical. In other words, employees receive an identical bias from any evaluator that is assigned by the company/organization implying that there is no need to adjust the employee's score with respect to an evaluators' effect. On the other hand, if the hypothesis $\sigma_2^2 \neq 0$, then different evaluators have a different levels of leniency/severity that they employ when judging an employee's performance, and thus the employee's score should be adjusted to account for this effect.

In a fixed effects model, my main interest will focus on whether the β_j 's are identical for all $j=1, 2, \dots, 70$. Such a model is known as a two-way mixed effect [12], [23]. If the data supports the following hypothesis $H_0: \beta_1 = \beta_2 = \dots = \beta_{70}$ then the employees will be received an identical bias from all the 70 evaluators so that there will be no need to adjust the employee's score for this evaluator's effect.

An important component of this model is a measure of its reliability sometimes called an intra-class correlation (ICC) coefficient, ρ , can be defined as the proportion of the total variation of the score that can be attributed to the true performance score.

The value of α_i , the employee's performance variation, can be estimated by a technique which is known as Best

Linear Unbiased Estimation (BLUE). BLUE is a class of statistical tools that has some desirable properties [7], [20]. The term "Best" in the acronym BLUE is used to describe the property that, from the available data on an employee, its estimated true performance will be as error-free as possible. The term 'linear' simply means the data has been adjusted to some other scale which has linear relationship between the variables. 'Unbiased' means that, on the average, the estimated true performance calculated from the model will be the same as the employee's true performance. 'Estimation' refers to the task at hand: trying to estimate the true performance. Once a BLUE has been obtained for each one of the employee based parameters, a hypothesis test can be constructed by noting that the standardized BLUE's are distributed as a student's t -distribution with degrees of freedom equal to the denominators degree of freedom (ddf). One can then pinpoint the i^{th} employee as being a significantly highest/least performer, if the standardized BLUE is greater than $t_{(1-\frac{\alpha}{2}, \text{ddf})}$ where $t_{(1-\frac{\alpha}{2}, \text{ddf})}$ is the lower $1-\alpha/2$, level of student's t distribution with denominator degrees of freedom ddf. For exceptionally highest performance, the estimate will be positive valued and for least performers it will be negative valued.

Model diagnostics also form an important part of tactical modeling. There are some formal and informal procedures that can be used to detect outliers, influential points and specific departures from underlying assumption in the linear mixed models [28]. These procedures will also be employed in this paper.

4. Results and Discussions

4.1. Without an Adjustment for the Evaluators' Effect

One can perform analysis without adjusting for the evaluators' effect, by simply using the average score that has been assigned by all the evaluators to a given employee. Using this approach, the highest and least performers are presented in Table 2.

Table 2. The least and highest performance employees using the mean performance scores.

Least performance		Highest performance	
Employee	Mean score	Employee	Mean score
87	9.00	78	23.00
115	10.00	188	23.00
19	11.00	86	22.30
85	12.00	89	22.30
91	12.00	1	22.00
191	12.00	212	22.00
108	13.00	189	22.00
153	13.00	119	21.80
176	13.00	37	21.50
178	13.00	34	21.00
192	13.00	64	21.00
210	13.00	199	21.00
47	13.50	93	20.60
150	13.50	29	20.30

4.2. Adjusted Model 1: Including an Evaluators' Effect as a Fixed Effect

Result for the evaluators' fixed effects model is given in table 3. The evaluator row of table 3 is testing whether the evaluators' effect parameter estimator that we have obtained are significantly different from zero. The very small p-value that we have obtained ($P=0.0001$ which is less than for all level of significance, for $\alpha=0.05, 0.01$) indicates that the hypothesis $H_0: \beta_1 = \beta_2 = \dots = \beta_{70} = 0$ can be rejected.

This clearly shows that the existence of an evaluator's bias in the scores given to different employees of the firm/organization. The variance parameter estimated for σ_1^2 that is given in Table 3 indicates that there is also Variability in the performance between employees that are statistically significant and therefore needs to be accounted for. From table 4 and using Best Linear Unbiased Estimators (BLUE), 73% of the total variance associated with the employees' score is attributable to the true performance score variability of the employees' σ_1^2 .

Table 3. Test on fixed and random effects' significance from fitting evaluator's effect as fixed

Fixed effect	Numerator df	Denominator df	F	Pr>F/
Evaluator: σ_2^2	69	59	5.71	0.0001
Random Components Variance parameter estimate	Estimate	Standard error	Z	Pr> Z
Employee: σ_1^2	2.69	0.411	6.56	0.0001
Error: σ_0^2	0.98	0.147	6.69	0.0001
Overall mean parameter estimate	Estimate	Standard error	T	Pr> t
Overall mean: μ	19.39	1.54	12.56	0.0001

Table 4 provides a ranking of employees based on the BLUEs that have been obtained for α_i . The results need to be interpreted as a continuum where large negative values indicate poor performance and large positive value indicate an excellent performance. An estimate for each employee's true performance score can then be obtained by adding the appropriate BLUE score that has been given in Table 4 for a given employee to the overall mean estimate of 19.39 that

has been given in Table 3. Unlike the results in Table 2, the results in Table 4 account for the bias from evaluators and adjust the employees' score for this evaluator's effect. Besides the adjustment for the evaluators' bias, Table 4 accounts for the variability of the employee score. For instance, employee 100 was not listed as one of the least performer employee in Table 2 but is listed as the second least performer employee in Table 4

Table 4. The least and best performance employees from fitting evaluators' effect as fixed

least performer employees				highest performer employees			
Employee No.	BLUE	SE	Pr> t	Employee No	BLUE	SE	Pr> t
87	-3.975	0.942	0.0001	155	4.082	0.942	0.0001
100	-3.279	0.826	0.0002	188	3.853	0.939	0.0001
115	-3.099	0.934	0.0016	37	3.723	0.728	0.0004
126	-2.930	0.958	0.0034	35	3.493	0.934	0.0015
74	-2.930	1.001	0.0049	189	3.12	0.939	0.0025
176	-2.914	0.895	0.0019	78	3.053	0.967	0.0034
178	-2.893	0.899	0.0021	121	2.93	0.958	0.0018
158	-2.823	0.824	0.0011	89	2.884	0.883	0.0045
135	-2.730	0.761	0.0007	75	2.761	0.934	0.0155
6	-2.451	0.658	0.0004	199	2.747	1.102	0.003
19	-2.366	0.934	0.0140	200	2.653	0.858	0.0213
138	-2.271	0.729	0.0028	86	2.284	0.966	0.0017
195	-2.181	0.895	0.0179	64	2.275	0.692	0.0167
132	-2.161	0.899	0.0194	183	2.261	0.918	0.0051
173	-2.161	0.899	0.0194	72	1.833	0.631	0.0128
38	-2.050	0.688	0.0042	141	1.72	0.67	0.0147
184	-1.936	0.697	0.0073	194	1.72	0.684	0.0138
187	-1.720	0.599	0.0057	36	1.395	0.55	
41	-1.611	0.537	0.0040				

When we scrutinize the evaluation report of employee 100, we see that employee 100 was evaluated by two evaluators (evaluators 32 and 58, with a score of 15.9 and 12.5 respectively). However, these two evaluators evaluated other employees; for example evaluator 32 evaluated eight employees and gave them scores of 18.2,

21.8, 16.6, 22.3, 20.6, 17, 19.6 and 15.9 respectively and evaluator 58 evaluated two employees with scores of 20 and 12.5 respectively. From the two evaluators we note that the score of employee 100 is the least. Moreover, by tracing back to determine how evaluators 32 and 58 evaluated other employees relative to the other evaluators, we notes

that, on average, evaluators 32 and 58 tended to be more lenient. With all these considerations in the model, the predicted performance score for employee 100 then becomes a significantly negative score, as given in Table 4. But the crude average score of employee 100, 14.2, would not place this employee among the least performance employee.

Likewise by adjusting for evaluators' effect employee 155 becomes one of the highest performers, as shown in Table 4, whereas this employee was not listed as a highest performer in Table 2.

The normality assumption and the goodness of fit for data used to conduct this paper, is also supported by W-statistic which indicates that no recognizable outlier in the data. The application of a more formal test [27] also did not record the maximum absolute studentised residual as being an outlier. The W-statistics which is an adaption of Shapiro and [26] normality test to a linear mixed model [28]. In particular, the following result was recorded ($W=0.9777$ for which favors the normal distribution).

Focusing on those observations, that could be potential outliers for our study, it was found that observation numbers 123 and 246 were the most influential observations. When these observations were removed, however, no significant change in the parameter estimates or goodness of fit of the resulting model was recorded. Nevertheless, because we are dealing with people who we may want to incentives it could be argued that one would like to examine these two outliers more carefully.

Observation number 123 contains a score of 15 for employee 72 that has been given by evaluator 27. This same employee was also Evaluated by three other people (namely, evaluator 21, 37 and 58) who gave that employee the following respective scores (22, 18, and 19.6). It should be noted that evaluator 27 also had to evaluate nine other employees and the score of employees 72 was the lowest given by evaluator 27. Evaluator 21, 37 and 58 put employee 72 as their 3rd, 3rd and 2nd highest performance employees, respectively. Case number 246 deals with employee 155 who was evaluated by a single person (evaluator 35), and was given a score of 20.

It should be mentioned that evaluator 35 also had to evaluate seven other employees (33, 73, 87, 155, 162, 194,

and 202) and gave them the following respective scores (13, 15, 9, 20, 12, 18 and 15). In terms of the evaluation that these seven employees received from other people, the score of evaluator 35 was found to be the lowest for five of these employees and the second lowest for another one of these employees. Because of this oblivious downward bias in the evaluation record of evaluator 35 when an adjustment is being made to employee 155's score, the predicted performance score for employee 155 then becomes very large as reflected in table 4.

4.3. Including an Evaluators' Effect as a Random Effect

Maximum likelihood estimates for the model parameter and the associated tests of significance are presented in table 5.

The result indicates that the evaluator and employee effects are significant. Employing our formula outlier testing procedure does not label any observation as being an outlier. None of the observations appear to be separated from the bulk of other observations. The summary statistic ($W = 0.978$), also favors a normality assumptions ($P=0.0860$).

Table 5. The parameters estimate from fitting evaluators' effect as a random effect

Random effect variance				
	Estimate	Standard error	Z	Pr> z
Evaluator	2.240	0.558	4.02	0.0001
Employee	2.443	0.484	5.04	0.0001
Error	1.903	0.314	6.05	0.0001
Overall mean				
	Estimate	Standard error	T	Pr> t
Overall mean	17.192	0.2274	75.57	0.0001

A prediction of the true performance of each employee shows that ten employee (see Table 6) can be regarded as performing exceptionally badly or well. For exceptionally good performers, note that the estimate will be positive valued and for bad performers the estimate will negative valued. Furthermore, the prediction of an employee's true performance is obtained by adding the estimate given in Table 6 to the overall mean that we obtained for the model.

Table 6. The least and highest performance employees when the evaluators' effect is treated as being a random

Least performance				highest Performance			
Employee	BLUE	SE	Pr> t	Employee	BLUE	SE	Pr> t
87	-4.049	0.965	0.0001	37	3.414	0.812	0.0001
100	-2.641	0.933	0.0063	89	3.09	0.943	0.0018
115	-2.595	1.078	0.0192	188	2.876	1.091	0.0107
38	-1.929	0.797	0.0187	78	2.701	1.108	0.0178
187	-1.773	0.716	0.0161	155	2.689	1.09	0.0165
				64	2.494	0.797	0.0027

All the Least and highest performance given in Table 6 were also identified as the least and highest performers given in table 4. The consistency of the employees'

performance and the overall variability in the harshness and leniency shown by the 70 evaluators, where the only role players in Table 6 results. But the role players for Table 4

results were the average leniency or harshness of the evaluators who evaluated the employees and the employees' performance. Since employees who were evaluated by fewer evaluators have a less consistent performance predictor, the majority of the least or highest performers who were evaluated by only one evaluator were the least favored to be listed from Table 4 in to Table 6. For instance, consider employees 37 and 86 from the top performer employees given in Table 4. Employee 37 was evaluated by three evaluators with a score of 21, 22 and 22. On the other hand, employee 86 was evaluated by a single rater with a score of 22.3. Employee 37 is a consistent performer and leads the top performers in Table 6, but not so employee 86. Since the evaluators' effects were considered as random effects, we obtain the BLUE of the realized evaluators' effect. An investigation of these estimates of the BLUEs of evaluators' effect showed the harshness or leniency displayed by evaluators in their judgments. Table 7 provides the extreme rankings of evaluators based on the BLUE's estimate of the evaluators' effect latent values: large negative values indicate a harsh evaluator and large positive values indicate a latent evaluator.

Table 7. Too harsh or too lenient evaluators

Evaluator	BLUEs	SE	DF	T	Pr> t
23	-2.577	0.575	59	-4.48	0.0001
48	-2.334	0.966	59	-2.42	0.188
71	-2.156	0.808	59	-2.67	0.0099
19	-2.010	0.570	59	-3.53	0.0008
35	-1.974	0.618	59	-3.19	0.0023
42	-1.486	0.628	59	-2.37	0.0213
32	1.610	0.662	59	2.43	0.0181
1	1.625	0.646	59	2.52	0.0146
3	1.661	0.690	59	2.41	0.0191
43	2.625	0.944	59	2.78	0.0073
31	3.664	0.611	59	5.99	0.0001

Evaluator 23, who evaluated thirteen employees and gave them the following scores 11, 19, 18, 12, 15, 16, 12, 15, 10, 14, 16, 14, and 13, can be viewed as being the most harsh evaluator. Similarity, Evaluator 31, who evaluated 6 employees and gave them the following scores 21, 21, 24, 22, 22, and 21, can be viewed as being the most lenient evaluator.

With regard to the existence of some possibly influential observations, observations number 69 and 297 were flagged in the analysis. Omitting both cases from the analysis did not substantially change the estimates that we obtained for the variance parameters or the overall goodness of fit of the model. It is interesting to note, however, that case 69 represents a score of 24 that was given to employee 39 by evaluator 43. This score was in fact the largest score that was given by any one evaluator to any one employee. The next highest score received by an employee was 16, which resulted in evaluator 43 being flagged an outlying evaluator in Table 7.

Case 297 refers to a score of 23 for employee 188, given by evaluator 40. This score is the second highest score that was given by evaluator in the entire employees' evaluation

process. Furthermore, this was the only score that employee 188 received.

In Table 2, results were based on the crude average score without any consideration of adjustment for the evaluators' effect. In Table 4 the employee performance predictor takes the average leniency/harshness of the associated evaluator in to consideration. In Table 5 the consistency of the employee in the evaluations, is taken in to account. What is evident from Tables 2, 4 and 6 is that the interest is in the true performance of the employee not in an average score based on a few measures/evaluators about the employee's performance.

The basic problem is that the observed value on the employee is not equal to the employee's true performance. How should we then estimate an employee's true performance latent value? The mixed model random effect links the evaluation to the true latent value. The estimate of the employee-true performance latent value is typically the BLUE estimate. As the number of measures on an employee gets larger, the BLUE estimate becomes consistent and approaches the employee's true performance latent value. The results in Table 4 and 6 are sufficiently convincing to use the BLUE estimates in employees' appraisal routine practice by considering evaluators' effect as fixed or random.

5. Conclusions and Implication

Performance appraisal systems are essential for a company to run efficiently and productivity. With performance appraisal in place, employees can be given a sense of ownership and responsibility with regard to the duties that they perform. The challenge is to know how best to adjust a given measure of an employee's performance so that it is not unduly influenced by evaluators tendency to make private and highly subjective assessments. Using a simple average of scores from a set of evaluators will not adjust for any hidden subjectivity that may reside in that specific group of rates. Because different employees are being assessed by different evaluators, a subjective bias may be introduced in to the rating of one employee when compared with that of another employee. This paper has sought to address this problem.

The linear mixed model that has been applied in this study allows for some flexibility with regard to whether one wants to view an evaluators' effect as being a fixed or random effect. An evaluator effect can be treated as being fixed if the evaluators are being selected by the company with the purpose of comparing one evaluator with another. On the other hand, the evaluator's effect can be treated as being random if we want to make statements about the variations in the overall population from which our evaluators are being drawn. Because we are interested in effects that, we believe, are common to all individuals and also effects that are different among individuals, a mixed effects model can be used to capture both these features. The mixed model provides estimates (BLUEs) of each

employee's true performance which can then be subjected to a formal test to identify those employees who, statistically, are significantly highest or least performers in the company.

The model's diagnostics tools that I have used help to provide some reassurance that the model is not being contradicted by the data that we are observing or being unduly influenced by particular characteristics of the data. The results of this paper have consistency shown that unless the same evaluators are evaluating all employees, there are considerable evaluators' effect which cannot be simply ignored in any employees' performance appraisal.

End note: The name of the company in which this survey has been conducted could not be disclosed for anonymity reasons.

Acknowledgment

The author is grateful to those individuals in a company who gave detail information and management group who gave attention for evaluating employees at those organization.

The author also grateful to those individuals who read the first draft carefully and gave their constructive comments.

Finally the author also grateful to the reviewer who gave highly important comments.

References

- [1] H. Aguinis, & C.A. pierce. Enhancing the relevance of organizational behavior by embracing performance management research. *Journal of Organizational Behavior*, 29: 139-145 (2008).
- [2] H.I. Braun, Understanding score reliability: experiments in calibrating essay readers. *Journal of Educational Statistics*, vol. 13: pp.1-18, 1988
- [3] D.N. Gruijter, D.N. Two simple models for rater effects. *Applied psychological measurement*, vol.8: pp. 213-218, 1984
- [4] G.R. Ferris, T.P. Munyon, k. Basik & M.R. Buckley. The performance evaluation context: social, emotional, cognitive, political, and relationship components. *Human resource management, Review*, vol.18: pp.146-163. 2008
- [5] J.L. Fleiss. *Design and analysis of clinical experiments*: New York: John Wiley & Sons, 1986
- [6] R.C. Grote. *The complete guide to performance appraisal*. New York, Amacom, AMA's book publishing division. 1996
- [7] D.A. Harville. BLUE (Best Linear Unbiased Estimation) and beyond. In *advances in statistical methods for genetic improvement of livestock*, pp. 239-276. New York: Springer-Verlag, 1990
- [8] B. Hoffman & D.J. Woehr. Disentangling the meaning of multisource feedback source and dimension factors. *Personnel psychology*, vol. 62, pp. 735-765, 2009.
- [9] B.J. Hoffman, C. Lance, B. Bynum & B. Gentry. Rater source effects are alive and well after all. *Personnel psychology*, vol.63, pp.119-151, 2010.
- [10] W.M. Houston, M.R. Ranymond & J.C. Sec, Adjustments for rater effects, *Applied Psychological measurement*, vol. 15, pp.409-421. 1991.
- [11] J.N.Kondrasuk, So what would an ideal performance appraisal look like? *Journal of Applied Business and Economics* vol.12 (1), pp.57-71, 2011
- [12] T.D. Little, K.U. Schnabel & J. Baumert. *Modeling longitudinal and multilevel data*. London: Lawrence Erlbaum Associates Publishers, 2000
- [13] A. Marganave & R.Gordan. *The complete idiots' guide to performance appraisals*. NewYork:Alpha Books/Macmillan.2001
- [14] Mc. Culloch, C.E. Searle, S.R. & G. Casella, *Variance components*. New York, John Wiley. 1996
- [15] Mc. Culloch, C.E.Seale & J.M. Neuhaus. *Generalized linear and mixed models*, 2nd ed. New York, John Wiley. 2008
- [16] B. Ounfiwira, B.Bourdage, & K.Lee. Rater personality and performance dimension weighting in making overall performance judgments. *Journal of business and Psychology*, vol.25,pp. 465-476, 2010.
- [17] E.D. Pulakos. The development of training programs to increase accuracy on different rating forms. *Organizational Behavior and Human Decision processes*, vol.38:pp. 76-91.
- [18] G.K. Robinson. The estimation of random effects, *Statistical Science*, vol.6, 1986.
- [19] M.Russill. Summarizing change in test scores: shortcoming of three common methods. *Practical Assessment, Research & Evaluation*, vol. 7 (5). Available at: <http://pareonline.net/getvn.asp?v=7&n=5> [Accessed 2009-09-16]. 2000
- [20] SAS institute. *SAS Technical Report p-229*. Carey (North Carolina): SAS institute Inc. 1992
- [21] S. Saxena. Performance management system. *Global Journal of Management and Business Research*, 10(5):27-30. 2010
- [22] S.S Shapiro, and M.B. Wilk.. An analysis of variance tests for normality (Complete Samples). *Biometrika*, vol.52: pp.591-611. 1965
- [23] A. Skrondal, A. & S. Rabe-hesketh, *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. London: Chapman and Halls. 2004
- [24] K.L. Ufferslev & L.M. Sulsky, Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology*, vol.93: pp.711-719. 2008
- [25] D.J.Woehr, M.K. Sheehan & W. Bennett. Assessing measurement equivalence across ratings sources: a multitrait-multirater approach. *Journal of applied Psychology*, vol.90, pp.592-600, 2005
- [26] E.W. Wolff. Identifying rater effects using latent trait models. *Psychology Science*, vol.46, pp.35-51, 2004

- [27] T. Zewotir. Influence diagnostics in mixed models. PhD thesis: University of Witwatersrand. 2001
- [28] T. Zewotir & J.S. Galpin. The behavior of normality under non-normality for mixed models. South African Statistical Journal, vol. 38, pp.115-138, 2004