Zambezia (1999), XXVI (i).

ASSESSING TEACHER PERFORMANCE: A COMPARISON OF SELF- AND SUPERVISOR RATINGS ON LENIENCY, HALO AND RESTRICTION OF RANGE ERRORS

T. J. NHUNDU

University College of Distance Education

Abstract

Self- and supervisor ratings of the performance effectiveness of teachers on 30 teaching and teaching-related tasks were obtained and compared to determine the potential usefulness of self-ratings. The study also compared self- and supervisor ratings to determine areas of agreement in perceived performance effectiveness of teachers and whether there was any correlation between the rating scores from the two sources. It was found that supervisor rating scores were more inflated and that supervisors tended to rate teachers globally instead of looking at specific performance tasks. However, a significant Spearman r (r=0.97) obtained from rank-ordered mean scores of the two groups showed that teachers and supervisors held similar perceptions over areas of lesser and greater performance effectiveness. Finally, Pearson correlation coefficients computed to determine the comparability of the rating scores of the two groups were statistically significant, indicating that both groups were measuring the same performance behaviours.

IN EDUCATION THE question of who should evaluate teacher performance is not as much an issue as are the purposes of evaluation or what should be evaluated and how it should be evaluated. There is greater concern over methodological issues than over key players in the evaluation process. It has almost become axiomatic in schools that teacher evaluation is carried out by supervisors and administrator-supervisors only and not peers and students and, least of all, supervisees themselves. There is, therefore, virtual dependence on teacher performance profiles provided by immediate, local, district, regional and central office personnel, contrary to emergent research findings which raise serious questions on the utility of continued dependence on supervisor appraisals as the only teacher evaluation approach in schools (Nhundu, 1992).

The question of who evaluates merits serious consideration because of advantages and disadvantages associated with a given evaluator, group or combination of evaluators. For example, the use of administrator evaluators in assessing teacher performance is likely to induce fear in the evaluatee due to perceptual dilemma resulting from contradictory bureaucratic and professional expectations inherent in administrative

and supervisory roles which reside in the same person. It is difficult to completely allay a supervisee's fears associated with the bureaucratic position occupied by the supervisor whose dual authority bases bring to the supervisory process a threatening atmosphere. Teachers, therefore, see the role of supervisors who also occupy administrative positions as not directly related to the improvement of instruction but associate them with supervision for administrative decisions.

Traditional performance ratings using superiors may, therefore, not be the best teacher evaluation method. On the contrary, teachers view self-ratings as the most appropriate evaluation method compared with supervisor/administrator and peer evaluation which they ranked second and third, respectively (Stark and Lowther, 1984, 97). Research findings also show that teachers are not happy with traditional assessment practices (Mclaughlin, 1984; Reavis, 1978; and Wolf, 1973). Hence, Levin (1979) and Paulin (1980, 10) have found that teachers, individually or through their professional organizations, have expressed unwillingness to be evaluated, especially when they do not trust the evaluator's expertise and also when they are not represented in both the design and implementation of the evaluation.

Self-evaluations, on the other hand, have the greatest potential of producing changes in teaching practices because they provide teachers with the rare opportunity to reflect on their teaching and modify accordingly. Johnston (cited in Balzer, 1973) compared the effects of traditional and self-evaluation practices on behaviour modification and found that self-ratings showed greater potential in changing teaching behaviour than traditional approaches. This finding is also supported by Natriello (1977) who cites similar evidence from his studies with the US armed forces.

Unfortunately, when self-ratings are obtained on a compared-to-others basis, research shows that there is greater leniency, less halo error and less variability (restriction of range error) on the part of self-evaluations (Ash, 1980; Heneman, 1974; Hubert and Dueck, 1985; Johnston and Sackeney, 1982; Klimoski and London, 1974; Levin, 1980; Meyer, 1980; and Thornton, 1968 and 1980). Restriction of range error occurs when one set of paired corresponding standard deviations obtained from independent performance ratings of two rating sources is significantly smaller. Smaller standard deviations indicate a relatively narrow range in the distribution or spread of the rating scores from which they are based.

On the other hand, leniency error arises when inflated performance effectiveness scores from one rating source are significantly different from the ratings obtained from other sources on the performance of the same group of ratees. This, in turn, often presents a measurement problem

because it narrows the range (spread) of possible performance ratings of ratees. According to Holzbach (1978, 579),

> Leniency errors present a measurement problem to the extent that restriction range on the performance ratings limits the magnitude of the potential relationship between ratings of performance and other variables of interest.

In addition, rating scores that are inflated may send incorrect signals to the ratee since they misrepresent a person's performance effectiveness.

Finally, halo error is a type of rater bias which arises when a rater fails to distinguish specific job dimensions in the appraisal process and, instead, employs global assessment. The relative incidence of halo effect is obtained when the magnitude of intercorrelations for supervisor ratings are compared with those for self-appraisals which are independently obtained using the same performance rating instrument. Hence, a rating source which produces larger intercorrelations indicates greater halo effect. In this connection, previous research has shown intercorrelations for supervisor ratings to be consistently higher than corresponding intercorrelations for self-ratings, indicating greater halo error for supervisor ratings (Heneman, 1974, 642).

Whilst a number of studies might have reported higher leniency errors with self-ratings, Heneman (1974, 642) and Miner (1968) argue that high leniency errors of self-ratings may be attributable to the purpose to which the ratings will be put. They have concluded that when the purposes of self-ratings are not administrative but research only, self-ratings may not be as inflated as reported in other studies. Subsequent studies by Holzbach (1978) and Nhundu (1992) concur with Heneman and Miner.

Nevertheless supervisor ratings remain the dominant methodology for evaluating teachers in spite of continued teacher discontent with this method. Furthermore, the assessment of teacher performance using supervisor-ratings is often too sporadic, and supervisory visits are too few and far apart in their frequency that they may not have any meaningful effect in the modification of teaching behaviour. Such supervisory practices are also often superficial because they are too broadly focussed and all-encompassing to stimulate teacher change and growth. In addition, the traditional supervisor-teacher rating relationship often creates insecurity and induces fear in the supervisee. According to Ness (1980, 405)

> There is an assumption, both implied and stated, that the authority to evaluate personnel carries with it fear of being judged, and this fear stands in the way of helping teachers . . . Little or no growth occurs as a result of a formal observation and instruction does not improve as a result of summative evaluations.

Such criticisms of traditional supervisor ratings raise serious questions over the potential usefulness of teacher performance evaluations compared with self-ratings. The merits of self-ratings are notable in their effect to bring about changes in teaching practices. In addition, self-appraisals may also be associated with teacher self-esteem and higher productivity (Meyer, 1980). Accordingly, if the crucial question in performance evaluation is accepted as the extent to which it produces changes in teaching practices and results in teacher growth, then self-appraisals hold the greatest potential in this regard. Thus, in view of lack of consensus concerning current status of self-ratings compared with supervisor ratings, more research should be undertaken to assess the potential usefulness of self-ratings in terms of their relative leniency, halo effects and variability.

In short, the main purposes of this study were to assess the potential usefulness of self-ratings as an alternative evaluation method by (a) obtaining self- and supervisor performance rating scores on selected teaching and teaching-related behaviours, and (b) assessing areas of greater and lesser performance effectiveness of teachers using self- and supervisor ratings, (c) determining and comparing self- and supervisor ratings in terms of their leniency, halo effects and variability (range error), and (d) assessing the rating scores of teachers and supervisors to determine the level of agreement in their selection and ranking of performance effectiveness dimensions.

RESEARCH METHOD

Supervisors were asked to assess the performance of their teachers using a thirty-item performance assessment questionnaire graduated on a five-point rating scale ranging from 1 = poor to 5 = exceptional. The performance assessment questionnaire was sent to a large randomly selected sample of teachers and their corresponding supervisors as part of a larger study on job satisfaction (Nhundu, 1994). The questionnaires for teachers and supervisors contained identical performance scales. Teachers were asked to rate their performance in teaching and teaching related tasks while supervisors rated these teachers on the same scales. Participation in the study was voluntary and the participants were assured of the confidentiality of their responses. Respondents were further informed that their responses were to be used for research purposes only.

However, participation in the study was done pairwise, using a dependent random sample comprising 229 teachers and their corresponding supervisors (N=229). Only certificated teachers with more than three years of teaching experience took part in this study. It was felt

that since untrained teachers were not professionals, their understanding of the teaching profession would be limited. Teachers who had not persisted beyond the heavy attrition period of three years, on the other hand, were considered less experienced and, thus, assumed to be unfamiliar with the rating scales and the measurement constructs and their expectations. According to Thornton (1980, 450), objectivity of self-assessments depends, in part, on the accuracy with which rating scales are interpreted. Thus, familiarity with rating scales leads to clearer understanding of the meaning of concepts being measured which, in turn, results in accurate interpretations and objective measurements.

RESULTS

Comparison of Areas of Greater and Lesser Performance Effectiveness

Teachers' Ratings of the Effectiveness of their Performance

Teachers were requested to rate their performance on selected teaching and teaching-related tasks using a five-point Likert-type rating scale ranging from 1 (poor), 2 (fair), 3 (good), 4 (very good) and 5 (exceptional). The thirty tasks on which they rated the effectiveness of their performance were further classified into four broad job dimensions, viz; Curriculum and Instruction (CI), Human Relations (HR), Personal Development (PD) and School-Community Relations (SCR). Means and standard deviations were computed for each of the thirty tasks. Their responses were rank-ordered to reveal areas of greater and lesser effectiveness.

Table 1 below lists the top ten tasks which were rated by teachers as their areas of greater performance effectiveness. The overall mean score for the ten top tasks was 3.52 which is greater than the theoretical mean score of 3.00 (assuming normal distribution of responses). All the top ten tasks had mean scores above the theoretical mean.

Of the top rated ten tasks appearing in Table 1, nine were classified as "Human Relations", and one as "Curriculum and Instruction". The apparent preponderance of human relations tasks in the top rated ten tasks clearly indicates that teachers in the research sample were most concerned with idiographic dimensions of their job than with the nomothetic aspects of teaching. The teachers' performance effectiveness in idiographic-related tasks was generally superior compared with other tasks. These results thus suggest that teachers in the sample valued more and performed better where their relationships with superiors, fellow teachers and students were concerned than in other aspects of their job.

Table 1
TOP TEN TASKS RANKED ACCORDING TO TEACHERS' RATINGS OF THE EFFECTIVENESS OF THEIR PERFORMANCE ON SELECTED TEACHING AND TEACHING-RELATED TASKS (N=229)

Rank	Category*	Performance Task	Mean	sd
1	HR	Maintaining good rapport with colleagues	3.74	0.93
2	CI	Classroom management and control	3.64	0.83
3	HR	Ability to make friendship with colleagues	3.61	0.95
4	HR	Maintaining good rapport with superiors	3.60	0.98
5	HR	Assisting in extra-curricular activities	3.49	1.00
6	HR	Participating in staff meetings	3.46	0.97
7	HR	Consistence and fairness with students	3.45	0.88
8	HR	Providing good leadership	3.44	1.12
9	HR	Reflecting and acting upon supervisory advice	3.41	0.87
10	HR	Cooperating with colleagues in lesson planning	3.40	0.92

**Category: HR = Human Relations; CI = Curriculum and Instruction*

The above finding further highlights the importance of human relations as a possible source of job satisfaction among teachers in Zimbabwe. According to Pigge and Lovett (1985) and Siegel and Bowen (1971) cited by Nhundu (1992), job satisfaction is both a result of, and dependent on, good performance. Accordingly, job satisfaction for teachers in this sample would more likely derive from human relations aspects of teaching where their perceived performance effectiveness was greatest compared with performance in other areas of their job.

Table 2 which lists the lowest rated ten performance tasks shows that four of these tasks belonged to "Curriculum and Instruction", three were on "School-Community Relations", one was on "Human Relations", and two were on "Personal Development" job dimensions. According to Table 2, all but three of the mean scores for the least rated tasks had values above the theoretical mean score of 3.0. The least rated task had a mean score of 2.60 and the highest rated received a performance rating score of 3.24, while the overall mean score for the lowest rated ten tasks was 3.01. The overall rating mean score for the ten bottom ratings shows that teachers in the sample rated their performance on these tasks as

"good". But when compared with the corresponding overall mean score for the top ten tasks of 3.52, the performance of teachers on the ten bottom tasks is substantially inferior.

Table 2
BOTTOM TEN TASKS RANKED ACCORDING TO TEACHERS' RATINGS OF THEIR PERFORMANCE EFFECTIVENESS ON SELECTED TEACHING AND TEACHING-RELATED TASKS (N=229)

Rank	Category*	Performance Task	Mean	sd
21	CI	Keeping accurate records	3.24	0.94
22	HR	Showing empathy for students	3.18	0.87
23	CI	Preparation of long and short term plans	3.18	0.87
24	CI	Appropriateness of lesson introduction and closure	3.16	0.81
25	CI	Responding to students' needs, aptitudes and learning styles	3.14	0.84
26	PD	Ingenuity and innovativeness	3.07	0.84
27	PD	Student counselling	3.01	0.95
28	SCR	Encouraging parental involvement in student learning	2.91	1.11
29	SCR	Participation in community activities	2.63	1.25
30	SCR	Holding parental conferences	2.60	1.11

**Category: CI = Curriculum and Instruction; PD = Personal Development; SCR = School-Community Relations; HR = Human Relations*

The preeminence of "Curriculum and Instruction" followed with "School-Community Relations" tasks among the ten bottom rated tasks shows that teachers in the sample considered themselves relatively less competent in carrying out the tasks which are central to the teaching profession, that is, curriculum and instruction and, in particular, issues concerning recent government policy towards greater community and parental involvement in local school governance. The fact that a preponderance (70%) of the ten least rated performance tasks belonged to these two job dimensions may suggest that teachers perceived their performance in these areas to be relatively weaker. Hence, in view of the centrality of curriculum and instruction issues to a school's mission and the emerging parental role in school-based decision making, the predominance of these two job facets in the ten bottom rated tasks should be considered from the perspective of the potential which

diminished teacher performance in these areas may have on teaching and educational standards.

From this finding, it also appears that concern for the human side of the school enterprise (where only one item appeared among the ten bottom rated tasks) is emphasized at the expense of pedagogy and pedagogy-related issues. This finding should be a cause for concern for policymakers, educationists and school administrators in Zimbabwe. Accordingly, Government's recent shift from quantitative expansion to qualitative improvement in primary and secondary education and enhanced local school governance should take cognisance of related research findings so that appropriate teacher training intervention programmes (including pre-service) can be designed to prepare, improve and strengthen teacher performance in curriculum and instruction-related areas. This finding also has important implications for in-service and other staff development programmes which seek to raise teaching competencies of practising teachers.

Supervisors' Ratings of the Effectiveness of Teachers

Supervisors were requested to assess the performance of teachers on the same dimensions, using a performance assessment questionnaire identical to that used by teachers. Their mean rating scores which now appear in Tables 3 and 4 below were rank ordered to determine areas of lesser and greater teacher performance effectiveness.

The overall mean score for supervisor ratings of the top rated ten tasks listed in Table 3 was 3.69 compared with 3.52 obtained for teacher ratings. Table 3 also shows that the areas of greater teacher performance (according to supervisor ratings) were also predominantly in human relations which accounted for seven of the ten top ranked tasks. A comparison of teacher ratings of their performance and the supervisors' ratings of the performance effectiveness of teachers which appears in Tables 1 and 3 respectively, shows a remarkably close agreement between the two independent ratings.

Firstly, there is general agreement that teacher performance effectiveness is greatest in the area of human relations. Tables 1 and 3 further show that all the seven human relations tasks rated highly by teachers were also rated in nearly the same rank order by supervisors. Finally, the first four tasks that received the highest performance effectiveness scores according to teachers' ratings are identical to those on the supervisors' list except that the mean performance scores for supervisors are slightly inflated compared with those for their supervisees.

Table 3
TOP TEN TASKS RANKED ACCORDING TO SUPERVISORS' RATINGS OF THE PERFORMANCE EFFECTIVENESS OF TEACHERS ON SELECTED TEACHING AND TEACHING-RELATED TASKS (N=229)

Rank	Category*	Performance Task	Mean	sd
1	HR	Maintaining good rapport with colleagues	3.92	0.90
2	CI	Classroom management and control	3.83	0.90
3	HR	Ability to make friendship with colleagues	3.77	0.94
4	HR	Providing good leadership	3.65	0.93
5	HR	Maintaining good rapport with superiors	3.64	0.93
6	HR	Assisting in extra curricular activities	3.63	1.07
7	HR	Consistence and fairness with students	3.63	0.88
8	HR	Participation in staff meetings	3.62	1.09
9	PD	Initiativeness	3.60	0.90
10	PD	Ability to make independent decisions	3.59	0.87

**Category: HR = Human Relations; PD = Personal Development; CI = Curriculum and Instruction*

The areas of least teacher performance effectiveness according to the supervisors' assessment appear in Table 4 below. Six of these tasks belong to 'Curriculum and Instruction', two to 'School-Community Relations' and one each to 'Human Relations' and 'Personal Development'.

The overall mean performance rating score for the ten bottom rated tasks computed from supervisor ratings was 3.20 compared with 3.01 obtained from the self-ratings by teachers. This indicates that while the overall performance mean rating scores for both sub-groups were above the theoretical mean score of 3.00, indicating that the two sub-groups rated the performance of teachers on the ten bottom rated tasks as good, both the overall and item-by-item rating scores of supervisors remained consistently inflated than those for self-ratings.

Furthermore, supervisors' assessment of the areas of least teacher performance effectiveness agrees with the assessment by teachers in six of the ten bottom tasks. There was also general agreement between teachers and their supervisors that the job dimension that was least performed by teachers was 'Curriculum and Instruction'. For teachers, four of the bottom ten tasks belonged to this job facet while supervisors' assessment identified six of the bottom ten tasks as belonging to the

Table 4

BOTTOM TEN TASKS RANKED ACCORDING TO SUPERVISORS' RATINGS OF THE PERFORMANCE EFFECTIVENESS OF TEACHERS ON SELECTED TEACHING AND TEACHING-RELATED TASKS (N=229)

Rank	Category*	Performance Task	Mean	sd
21	CI	Organising student learning activities	3.33	0.86
22	CI	Developing challenging teaching activities	3.32	0.92
23	CI	Appropriateness of lesson introduction and closure	3.31	0.89
24	CI	Preparation of long and short term plans	3.29	0.78
25	HR	Cooperating with colleagues in lesson planning	3.19	0.92
26	CI	Responding to students needs, aptitudes and learning styles	3.18	1.00
27	CI	Suitability of learning materials, illustrations, etc.	3.18	0.99
28	PD	Ingenuity and innovativeness	3.13	0.79
29	SCR	Encouraging parental involvement in student learning	3.04	1.09
30	SCR	Participation in community activities	2.62	1.24

*Category: HR = Human Relations; CI = Curriculum and Instruction; SCR = School-Community Relations; PD = Personal Development

same job facet. The next least performed job facet was 'School-Community Relations' whose tasks were the least performed of all the bottom ten least rated tasks. However, the ratings of supervisors remained consistently, but slightly, inflated compared with those for supervisors. Variations in mean ratings ranged from 0.02 to 0.08 between the least performed and best performed of the bottom ten least performed tasks, respectively.

Comparison of the Perceptions of Teachers and their Supervisors Concerning Relative Leniency, Restriction of Range and Halo Errors

Leniency: Means and standard deviations were computed for all 30 job dimensions and these were used to compare the relative leniency and range errors, respectively. The results of this analysis appear in Table 5 below. While these results show that 24 of 30 mean supervisor rating scores were larger than self-ratings, and that only one mean rating score

was the same for both groups, this however, does not allow for the rejection of the null hypothesis that there is no difference in the ratings of the two groups. To test the hypothesis, a Wilcoxon matched-pairs signed-rank test was computed. The test takes into account the magnitude and direction of the differences between paired mean rating scores of teachers and their supervisors obtained using the same rating scale.

The results of a two-tailed Wilcoxon matched-pairs signed-rank test analysis ($N=29$, $T=64$, $p>0.05$) showed that the ratings of supervisors were significantly higher than corresponding ratings of teachers, even at 0.01 level of significance. This finding shows that supervisor ratings had greater leniency error than self-ratings.

Table 5

MEANS AND STANDARD DEVIATIONS OF SELF AND SUPERVISOR RATINGS COMPUTED TO ASSESS RELATIVE LENIENCY AND RANGE ERRORS (N=458)

Performance Dimension	Supervisor Ratings		Self Ratings	
	mean	sd	mean	sd
Preparation of long and short term plans	3.29	0.78	3.18	0.87
Designing appropriate objectives	3.36	0.84	3.29	0.84
Organising students activities	3.33	0.86	3.33	0.78
Keeping accurate records	3.42	0.96	3.24	0.94
Cooperating with colleagues in lesson planning	3.19	1.04	3.40	0.92
Reinforcing students	3.55	0.96	3.34	0.89
Developing interesting and challenging learning activities	3.32	0.92	3.34	0.86
Responding to students' needs, aptitudes and learning styles	3.18	0.92	3.14	0.84
Using a variety of appropriate questioning techniques	3.37	0.91	3.29	0.90
Suitability of learning aids, illustrations, etc.	3.18	0.99	3.27	0.83
Appropriateness of lesson introduction and closure	3.31	0.89	3.16	0.81
Ingenuity and innovativeness	3.13	0.79	3.07	0.84
Showing empathy for students	3.39	0.88	3.18	0.87
Student counselling	3.43	1.00	3.01	0.95
Consistence and fairness with students	3.62	0.88	3.45	0.88

Table 5 (cont)

Performance Dimension	Supervisor Ratings mean	sd	Self Ratings mean	sd
Encouraging parental involvement in students' work	3.04	1.09	2.91	1.11
Holding parental conferences	2.39	1.15	2.60	1.11
Student supervision	3.52	0.85	3.39	1.01
Participation in community activities	2.62	1.24	2.63	1.25
Participation in staff meetings	3.62	1.09	3.46	0.97
Maintaining good rapport with colleagues	3.92	0.90	3.74	0.93
Initiativeness	3.60	0.90	3.25	0.82
Providing good leadership	3.65	0.93	3.44	1.12
Assisting in extra curricular activities	3.63	1.07	3.49	1.00
Reflecting and acting upon supervisory advice	3.44	0.89	3.41	0.87
Ability to make independent decisions	3.59	0.87	3.39	0.83
Maintaining good rapport with superiors	3.64	0.93	3.60	0.98
Developing own teaching approaches	3.54	0.91	3.39	0.90
Ability to make friendship with other teachers	3.77	0.94	3.61	0.95
Classroom management and control	3.83	0.90	3.64	0.83

Restriction of Range Error: Corresponding paired standard deviations used to assess restriction of range error (relative variability of rating scores) of supervisor and self-ratings also appear in Table 5. According to the results in Table 5, 19 of the 30 standard deviations were larger for supervisor ratings suggesting greater variability of supervisor ratings on these items. Of the remaining 11 standard deviations, nine were larger for self-ratings and two were the same for both groups. However, to determine whether these preponderantly larger supervisor ratings indicated overall significant differences between the ratings of teachers and supervisors, a two-tailed Wilcoxon matched-pairs signed-rank test was computed. The results of this analysis ($N=28$, $T=145$, $p>0.10$) failed to produce statistically significant differences between variances of supervisor ratings and corresponding self-rating variances, indicating that the observed differences might have been due to chance.

Halo Error: Intercorrelation matrices for performance dimensions for self-ratings and those for supervisor ratings were used to assess the relative halo error between performance ratings of the two groups. The monomethod-hetero-trait triangles (which were too large to include in this article) were used to investigate the incidence of halo error. The method involves comparing the sizes of intercorrelations obtained from more than one rating source (e.g. supervisors and teachers) based on rating scores independently obtained using the same rating instrument. When halo effect is measured using this method, a rating source which produces larger intercorrelations indicates greater halo effect.

According to intercorrelation matrices obtained for this analysis, there were 435 possible comparisons between self- and supervisor ratings (each triangle had 435 intercorrelations). However, only 425 comparisons were possible since 10 intercorrelations were the same for the two groups. Since the level of halo effect for this study was obtained by comparing the magnitude of intercorrelations for items obtained from the ratings of teachers and supervisors, a comparison of the 425 intercorrelations for the two groups showed that intercorrelations for supervisors were greater in 229 comparisons. However, to determine whether these preponderantly larger supervisor intercorrelation coefficients indicated overall significant differences between the ratings of teachers and supervisors, a two-tailed Wilcoxon matched-pairs signed-rank test was computed. The results of this analysis ($N=425$, $T=477$, $p>0.05$) show that intercorrelations for supervisors were significantly larger than those obtained for self-ratings, indicating that supervisor ratings had significantly greater halo error than self-ratings.

Overall Performance Assessment

An analysis of the perceptions of teachers and their supervisors on their ratings of all the 30 items on the questionnaire and also on the top and bottom ten tasks revealed that there was general agreement concerning their selection and ranking of the performance effectiveness of teachers. The overall mean performance score obtained from the ratings of teachers was 3.29 compared with a slightly inflated overall mean score of 3.47 for supervisors. A Spearman rank order correlation coefficient computed for rank-ordered means of teachers and supervisors on the 30 items produced a high rank order correlation coefficient (rho = 0.97), indicating a very strong positive relationship between the perceptions of the two groups. The high Spearman r obtained from rank-ordered means of the two groups further indicates that there was very high consistency in the rankings of teachers and supervisors. This also shows that the two groups held similar perceptions over areas of lesser and greater task performance

by the teachers. However, the mean scores for supervisors were generally slightly higher on all 30 tasks.

A two-tailed t-test analysis was run on the 30 items to determine whether the observed variances between the mean scores of teachers and supervisors were statistically significant.

Table 6

A T-TEST ANALYSIS OF THE PERCEPTIONS OF TEACHERS AND THEIR SUPERVISORS CONCERNING THE PERFORMANCE EFFECTIVENESS OF TEACHERS ON SELECTED TEACHING AND TEACHING-RELATED TASKS (N=458)

Category*	Performance Task	Mean Score Teacher	Supervisor	T-value	**p value
CI	Classroom management and control	3.83	3.64	1.99	0.048
PD	Initiativeness	3.60	3.25	3.63	0.000
PD	Ability to make independent decisions	3.59	3.39	2.05	0.041
CI	Reinforcing students	3.55	3.34	1.99	0.048
PD	Student counselling	3.43	3.01	3.83	0.000
HR	Showing empathy for students	3.39	3.18	2.10	0.037

*Category: CI = Curriculum and Instruction; HR = Human Relations; PD = Personal Development

**Probability value based on two-tailed test of significance

The results of the t-test analysis which appear in Table 6 above show that only six of the 30 job tasks produced statistically significant differences between the ratings of teachers and those of supervisors. Of the six areas where statistically significant differences emerged between the ratings of the two groups, three of the tasks were on 'Personal Development', two were in the area of 'Curriculum and Instruction' and one was on 'Human Relations'.

DISCUSSION

Supervisor ratings are commonly valued and used by school jurisdictions to acquire insight into teacher performance effectiveness and to assist them make administrative decisions because they are considered to be more objective and robust compared to other performance assessment methods. According to Holzbach (1978, 587), objectivity of supervisor ratings is attributable to the supervisors' wide experience and

responsibility in evaluating job performance as well as their familiarity and sensitivity to differentiate among specific job-related behaviours of individual supervisees. The findings of previous studies on rater bias in terms of leniency, halo and restriction of range errors were replicated in this study. Contrary to most previous research findings (Ash, 1980; Hubert and Dueck, 1985; Johnston and Sackeney, 1982; and Levin, 1980), this study showed that supervisor ratings produced greater leniency error than corresponding self-ratings. On variability and halo effects, the results of the present study are in agreement with previous research by Heneman (1974), Holzbach (1978) and Lawler (1967).

However, the results of this study on leniency provide more support to previous studies (Heneman, 1974; Miner, 1968 and Nhundu, 1992) which are at variance with the more prevalent findings of other studies which show that self-ratings have higher leniency error compared to counter position ratings. While supervisor ratings were consistently inflated, the differences in the ratings of the two groups were significantly different in only six of the 30 rating scales. Comparability in performance assessments between the two groups was determined using Spearman and Pearson correlation coefficients. The results also showed that self-ratings had significant correlations with supervisor ratings on identical performance tasks.

A possible explanation of these findings on leniency is that rater bias may be influenced by the rater's sensitivity and awareness of specific performance scales and job-related behaviours that contribute to measures of performance effectiveness. It is therefore possible that, because of the highly selective nature of the study sample which comprised of experienced teachers only, teachers in the sample were familiar with and had a clearer understanding of the rating scales and the job behaviours being measured. This would then make it possible for teachers to carry out more objective diagnostic assessments of their individual performance behaviours than supervisors who may have a more generalized global understanding of job-related behaviours of supervisees.

Additionally, the fact that the ratings were obtained under conditions where the findings were to be used for research purposes only may have contributed to more objective self-assessments as previously suggested by Heneman (1974). However, it is also important that research should seek to identify the sources of rater bias if it is going to contribute to meaningful improvement of performance practices. It is not enough for research to show that leniency error is attributable to rating sources without being able to identify the sources of leniency error. Accordingly, future research should seek to identify the sources of leniency attributable to rating sources. At the same time, future research should use more

rating sources such as peers, superiors, and students so that multiple comparative analyses can be carried out to establish the behaviour of self-ratings against these sources. The results of these studies will contribute towards a better understanding of current performance assessment practices, especially in universities where multiple rating sources are routinely used in evaluating lecturer performance.

The evidence obtained from this study clearly shows that although supervisor ratings exhibited greater leniency error than self-ratings, Pearson correlation coefficients computed to determine the comparability in performance ratings between supervisor and self-ratings revealed that their ratings were significantly correlated, contrary to previous research findings by Holzbach (1978). The current finding indicates that teachers and supervisors were measuring the same performance behaviours and also that they had a common understanding of the measuring scales used. Similarly, a very high Spearman rank order correlation coefficient obtained for rank-ordered mean scores for the two groups indicates that the teachers and their supervisors held similar perceptions over areas of greater and lesser teacher performance effectiveness.

Results on restriction of range error which produced larger (19 of 30 comparisons) variances for supervisor ratings failed to produce significant differences ($p<=0.10$, two-tailed Wilcoxon matched-pairs signed-rank test) indicating that the differences may have been due to chance instead. Since restriction of range error attributable to specific rating sources occurs only when the variances from different rating sources concerning the same ratee group are significantly different, the present finding suggests that the observed variability in scores for self- and supervisor ratings were, therefore, generally similar. This explanation is further supported by significant correlations which showed that the two groups shared similar perceptions on performance rating scales used in this study which, consequently, led to close agreement over areas of greater and lesser performance effectiveness.

The fact that the variability of the ratings of teachers and their supervisors was generally the same may further indicate that the two groups had clearer understanding of the concepts measured and the rating scales used, and also of the job performance behaviours of supervisees. A clearer understanding of the meaning of concepts being measured increases accuracy in the interpretation of rating scales (Thornton, 1980) which, in turn, might have helped narrow the range of variance between the two groups.

The finding on the incidence of halo error is consistent with previous research (Heneman, 1974; Holzbach, 1978 and Klimoski and London, 1974) which found that self-ratings contained less halo error than supervisor ratings. Since halo effect is a form of rater bias which results

when a rater assesses the performance of a ratee globally because of the rater's failure to differentiate among specific job performance behaviours, the results of this study suggest that supervisors tend to assess the performance of their supervisees globally. Their long experience and responsibility for routine subordinate performance evaluation may, inadvertently, influence supervisors to evaluate their subordinates globally without much reference to specific job performance behaviours. Thus, the findings of this study suggest that although supervisors may readily and globally identify a good teacher from a bad one because of their long years of experience in performance evaluation, data for training needs of teachers should be based on an assessment of their performance in specific job dimensions and not global performance. Accordingly, since self-ratings showed less halo error, indicating ability to discriminate among specific performance dimensions, teachers in the research sample tended to have greater awareness of areas of strengths and weaknesses in their performance than their supervisors who assessed them globally.

In conclusion, the findings of this study show that self-ratings have the greatest potential for providing the database from which a compendium of training needs of teachers can be compiled. Such a database provides scope for incorporation into pre-service teacher preparation programmes. The value of self-performance assessments highlighted in this study further suggests the need to provide teachers, during both pre-service and in-service training, with basic supervisory skills which will enhance their capacity for self-assessment. Equally important is the need to alert supervisors of the value of self-assessments and how these can inform and enhance the supervisory process.

Meanwhile, greater halo error associated with performance ratings of supervisors indicate that supervisor ratings tended to provide a global picture of teacher performance without identifying specific job-related behaviours that are of interest in the design of staff development programmes that seek improvement in teacher performance. Effective staff development programmes should address identifiable areas of deficiency in teacher performance; and such programmes can, therefore, benefit more from self-ratings. Hence, the presence of less halo error reported in this study may further explain why Balzer (1973) and Natriello (1977) concluded that self-ratings had greater potential of producing changes in teaching behaviours than supervisor ratings. What the present study has further found is that self-ratings are of potentially superior value to educational managers and educationists because they possess less leniency and halo errors compared to supervisor assessments. Whether self-ratings will retain less leniency and halo errors under conditions where the ratings are used for non-research purposes is a subject for further investigation. Similarly, future research should also

seek to identify the sources of leniency and halo effects so that performance counselling can be directed at the most needy areas for teachers.

References

Ash, R. A. (1980) 'Self-assessment of five types of typing abilities', *Personnel Psychology*, XXXIII, 273-282.

Balzer, L. (ed.). (1973) *A Review of Research on Teacher Behavior* (Columbus, Ohio, ERIC/SMEAC).

Heneman, H. G. (1980) 'Comparisons of self and supervisor ratings of managerial performance', *Journal of Applied Psychology*, LIX, 638-642.

Holzbach, R. L. (1978) 'Rater bias in performance ratings: Superior, self- and peer ratings', *Journal of Applied Psychology*, LXIII, (v), 579-588.

Hubert, B. D. and C. G. Dueck. (1985) 'On-the-job training of assistant principals in selected tasks in the Calgary School system', *Alberta Journal of Educational Research*, XXXI, (xxxi), 270-287.

Johnston, J. M. and L. E. Sackeney. (1982) *Principals' Classroom Supervisory Practices* (Saskatoon, University of Saskatchewan, College of Education).

Klimoski, R. J. and M. London. (1974) 'Role of the rater in performance appraisal', *Journal of Applied Psychology*, LIX, 445-451.

Lawler, E. E. (1967) 'The Multitrait-multirater approach to measuring managerial performance', *Journal of Applied Psychology*, LI. 369-381.

Levin, B. (1979) 'Teacher evaluation: A review of research', *Educational Leadership*, XXXVII, (iii), 240-245.

Levin, E. L. (1980) 'Introductory remarks for the symposium on organizational applications of self-appraisals and self-assessment: Another look', *Personnel Psychology*, XXXIII, 259-262.

Mclaughlin, M. W. (1984) 'Teacher evaluation and school improvement', *Teacher's College Record*, LXXXVI, (i), 193-207.

Meyer, H. H. (1980) 'Self-appraisal of performance', *Personnel Psychology*, XXXIII, 291-295.

Miner, J. B. (1968) 'Management appraisal: A capsule review and recent references', *Business Horizons*, I, 83-96.

Natriello, G. (1977) *A Summary of Recent Literature on the Evaluation of Principals, Teachers and Students* (ERIC Document Reproduction Service).

Ness, M. (1980) 'The administrator as instructional supervisor', *Educational Leadership*, XXXVII, (v), 404-406.

Nhundu, T. J. (1994) 'Facet and overall satisfaction with teaching and employment conditions of teachers in Zimbabwe', *Zimbabwe Journal of Educational Research*, VI, (ii), 153-194

— (1992) 'Job performance, role clarity, and satisfaction among teacher interns in the Edmonton Public School system', *The Alberta Journal of Educational Research*, XXXVII, (iv), 335-353.

Paulin, P. (1981) 'The Politics of Evaluation at the Local Level: A View Through Teachers' Perspectives', Paper presented at the Annual Meeting of the American Educational Research Association, 13-17 April 1981, Los Angeles, California.

Pigge, F. L. and M. T. Lovett. (1985) *Job Performance and Job Satisfaction of Beginning Teachers* (ERIC Document Reproduction Service).

Reavis, C. (1978) 'Clinical supervision: A review of research', *Educational Leadership*, XXXV, (vii), 580-584.

Stark, J. S. and M. A. Lowther (1984) 'Predictors of teachers' preferences concerning their evaluation', *Educational Administration Quarterly*, XX, (iv), 76-106.

Thornton, G. C. (1968) 'The relationship between supervisor and self-appraisal of executive performance', *Personnel Psychology*, XXI, 441-455.

—. (1980) 'Psychometric perceptions of self-appraisal of job performance', *Personnel Psychology*, XXXIII, 263-271.

Wolf, R. (1973) 'How teachers feel toward evaluation', in Ernest Haus, (ed.) *School Evaluation* (Berkley, California, McCutchan).