C E Center for Effective Organizations

Obtaining "Purer" or "Poorer" Criteria for Test Validation: An Empirical Test of the Statistical Control of Halo and Implications for Criteria Development

CEO Publication
T 87-19 (110)

Barry R. Nathan
University of Southern California

Nancy Tippins
Bell Atlantic Network Services, Inc.
Arlington, VA

May 1994

Paper under review at the Journal of Applied Psychology. Please do not cite without permission.

Obtaining "Purer" or "Poorer" Criteria for Test Validation: An Empirical Test of the Statistical Control of Halo and Implications for Criteria Development

CEO Publication
T 87-19 (110)

Barry R. Nathan
University of Southern California

Nancy Tippins
Bell Atlantic Network Services, Inc.
Arlington, VA

May 1994

Paper under review at the Journal of Applied Psychology. Please do not cite without permission.

Abstract

The assertion by Landy et al. (1980) that purer criteria for test validation would result from statistically controlling for halo in dimension ratings by partialing out a general impression rating was investigated. Data were analyzed from 294 clerical workers who participated in a validation study of 10 clerical ability tests, in which 17 duty and one overall performance ratings had been used as criteria. Statistically controlling for halo consistently result in poorer criteria. Residual rating scores typically resulted in nonsignificant correlations with predictor scores. Overall ratings gave, with few exceptions, the highest validation results of all test-rating combinations.

Obtaining "Purer" or "Poorer" Criteria for Test Validation: An Empirical Test of the Statistical Control of Halo and Implications for Criteria Development

In what has become a very controversial recommendation, Landy and his colleagues have suggested using a statistical approach to control for the presence of halo in performance rating (Landy, Vance, Barnes-Farrell, and Steele, 1980; Landy, Vance, and Barnes-Farrell, 1982). Specifically, they argued that an overall rating could be used as a surrogate variable for the halo effect and statistically partialed out of job dimension ratings. Criticisms of this technique have focused primarily on whether the technique is consistent with our current conceptualizations of the cognitive processes involved in performance rating (Feldman, 1986; Harvey, 1982; Hulin, 1982; Lance & Woehr, 1986; Landy et al., 1980, 1982; Mossholder & Giles, 1983; Murphy, 1982; Nathan & Lord, 1983). However, there has been neither discussion nor empirical testing of a practical implication of the technique that was suggested by Landy et al., namely that the residual scores that result from partialing out overall effectiveness ratings from each rating on specific performance items, "may represent 'purer' criteria for test validation and training evaluation purposes." (1980, p.505) In other words, the use of residual instead of observed performance scores as criteria in test validation, should result

in larger validity coefficients and greater probabilities that valid tests will be found to be statistically related to performance. On the other hand, if as its detractors suggest, the procedure removes true score variance as well as the systematic variance due to halo error, then statistically controlling for halo would result in poorer criteria for test validation and training evaluation and, therefore, lower validity coefficients.

The purpose of this study was to test the assertion that statistically controlling for halo would result in purer criteria by comparing test validities based on observed and residual performance scores. If Landy et al. (1980) are correct, the validity coefficients based on residual performance scores should be higher than those based on observed performance scores. Likewise, since the overall rating is assumed to be largely halo error, validity coefficients based on the overall performance rating should be lower than those from the residual performance ratings. On the other hand, the overall rating might not be the result of halo. Instead it could be a judgmental composite of what the rater believes is all of the relevant information necessary for making accurate ratings about a ratee. In this case, the overall rating would represent the most comprehensive sampling of the ratee's performance domain, and the most accurate judgment of ratee performance. Statistically removing this source of true score variance would result in lower validity coefficients

for the residual scores than for the observed performance ratings. Thus, partial correlation techniques could have the opposite effect of what is intended and result in less accurate performance criteria.

Method

Subjects and Ratings Forms

Ratees were 294 nonexempt clerical workers in a large petrochemical company; the raters were their immediate supervisors. Performance ratings had been made as part of a validation study of a clerical test battery. Ratings were made on 19 job analysis-based job duties and one overall performance ratings. Each duty (e.g., typing, ordering supplies and monitoring inventory, preparing written reports) also included a description of several specific tasks typically performed as part of the duty. Ratings were made on nine-point numerical rating scales with five verbal descriptors: poor, adequate, good, very good, and outstanding. There was also a space for raters to indicate if the duty was not applicable or could not be rated.

Predictors

For purposes of this study, test scores from nine short clerical ability tests were combined into three general abilities: clerical ability, consisting of combined standardized scores on tests of filing, coding, and checking; verbal ability, consisting of combined standardized scores on tests of vocabulary, grammar,

reading, and spelling; and numerical ability, consisting of combined standardized scores on a basic arithmetic test and a word problem test. In addition, results from a short mental ability test consisting of both verbal and quantitative questions are also reported.

Statistical Analyses

In a recent paper, Henik and Tzelgov (1985) have argued that statistically controlling for halo involves conceptualizing halo as a suppressor variable on the predictor-criterion relationship. In other words, halo is correlated with the criterion, dimension ratings but is unrelated to the predictor (tests). They suggest using a multiple regression approach, whereby a linear combination of the dimension and global ratings is regressed on a single predictor. This procedure allows one to calculate the part correlation between the residual scores of the dimension rating, i.e., the dimension scores with the effect of the overall rating removed, and each predictor score. If Landy et al. are correct, the part correlation coefficients from the residual scores should be larger than the zero-order correction coefficients between the predictors and uncorrected dimension ratings.

This type of design, in which measures of job performance are investigated by "validating" them against test scores, is not new to the literature. It was used by Bittner and Rundquist (1950) to compare rank-comparison ratings with measures of factory worker

output, by Rush (1953) to compare different criteria of sales performance, by Buchner (1959) to compare the predictability of ratings of submariner performance as a function of inter-rater agreement, and more recently, by Nathan and Alexander (1985a) in a validity generalization study comparing the predictability of different criteria of clerical performance.

Results

Because of the variety of clerical jobs included, some duties were not performed by all ratees and, therefore, were not rated. As a result, two of the 19 duty ratings, Performing Supervisory Functions ($\underline{N}$ = 23) and Performing Accounting Functions ($\underline{N}$ = 66), were dropped from our analyses because of small sample sizes. The remaining 17 duty ratings have sample sizes of at least 93.

Table 1 presents the intercorrelations among ratings. The observed correlations (above the diagonal) between each dimension rating and the overall rating are generally quite high, ranging from 0.52 to 0.78 with a mean of 0.59. In contrast, the correlation between individual dimensions, though all significant, are more moderate, ranging from 0.10 to 0.69. In other words, dimension ratings are far more like the overall rating than like each other. In contrast, the intercorrelations among dimensions with the overall rating partialled out (below the diagonal) are quite low and rarely significant, a condition generally interpreted as the absence of halo.

Insert Table 1 About Here

Table 2 presents the results of our regression analyses. The first row of the table shows that all four predictor variables correlate significantly with the overall performance rating. The rest of the table consists of the validation results for observed and residual score ratings of each job duty with each predictor. In every case, "statistically controlling for halo," reduced the validity coefficient. Only occasionally did the correlation based on the residual score remain significant; generally they were reduced to nearly zero or to a negative correlation.

Insert Table 2 About Here

The fact that controlling for the overall rating resulted in so many nonsignificant part correlations is disturbing; it implies that the use of dimension ratings adds little predictable variance beyond that due to a simple overall rating. To further investigate the relationship between dimension and overall ratings, the order of the ratings inserted into the hierarchical regression equation was reversed. This time, residual variance due to the overall rating beyond that due to the dimension rating was calculated. The incremental variance due to these two regressions is presented in Table 3.

Insert Table 3 About Here

As would be expected with such large dimension-overall correlations, the incremental $\underline{R}^2$s are almost all very small and nonsignificant. For clerical, verbal, and general mental abilities, duty ratings add little predictable variance beyond that due to the overall rating. One notable exception is the significant incremental $\underline{R}^2$ of typing duties with verbal ability. In contrast, for these same three predictors, overall performance ratings generally add predictable variance to that of the more specific duty ratings. This is very strong support for the argument that overall ratings are not error, but an important and predictable source of true variance.

The results for numerical ability are ambiguous. However, since in five of 17 duty ratings the incremental $\underline{R}^2$ is significant, there is some evidence that raters can discriminate among ratees within some job dimensions independent of overall impressions. This is particularly true in duties where numerical ability requirements would a priori seem most necessary, for example, Analyzing Data and Performing Arithmetical Calculations. However, this is still not support for statistical controlling for overall impression since, as Table 2 shows, the correlations based

on residual scores are still far smaller than those from observed scores.

To summarize, not only did the residual duty rating scores always result in lower validity coefficients than those resulting from the observed duty ratings, but they were with only three exceptions, lower than the global overall rating. Furthermore, it was the overall rating that added small but significant incremental variance to the dimensional ratings, not the reverse.

Discussion

The first and most obvious conclusion from these analyses is clear: do not statistically control for halo when conducting validation studies. The procedure removes almost all of the predictable variance in the criterion. Because these tests had been found to be valid predictors of performance prior to this analysis, and since it has been effectively argued that clerical tests are predictive of clerical performance across various situational variables (Pearlman, Schmidt, and Hunter, 1980, Schmidt, Hunter, and Pearlman, 1981), and criteria (Nathan and Alexander, 1985a), we can conclude that these tests measure abilities necessary for the job. Since removing the overall rating reduces the validity of these job-requisite abilities to zero, then the technique does in fact remove true variance as suggested by others (Feldman, 1986, Harvey, 1982; Hulin, 1982;

Lance & Woehr, 1986; Landy et al, 1980, 1982; Mossholder & Giles, 1983; Murphy, 1982; Nathan & Lord, 1983).

A second, and far more controversial conclusion is that, in general, dimensional ratings may be unnecessary when conducting validation studies, and that a simple overall effectiveness rating is sufficient. With few exceptions, the overall rating was larger than the observed job duty ratings. In fact, Table 3 shows that specific dimension ratings generally added little if any predictable variance to that attributable to the overall ratings. (In contrast, general ratings often added significant variance to that of the specific dimensions alone). In other words, failing to include ratings of overall performance would result in poorer estimates of tests' validities. Conversely, failing to include ratings on specific dimensions would have very little effect on the validity coefficients.

Given that these are valid tests, this means that raters are better at evaluating performance in general than performance at specific job duties. This conclusion is particularly troubling in light of the EEOC Guidelines and judicial decisions regarding performance appraisal. Both have strongly advocated the use of specific, behavioral dimensions and argued against the use of overall evaluations (Nathan and Cascio, 1986). If similar findings occur for other jobs, it would seem that companies are wasting time, effort, and money to develop elaborate and expensive

rating scales that meet with advocated, yet unsubstantiated technical and judicial standards for criterion development. (It should be noted that specific duty ratings may still be important for purposes other than validation criteria, such as employee feedback and development.)

What this says about how raters evaluate performance, i.e. the rating process, can be inferred only indirectly, but it does appear to support a categorization type rating process (Nathan and Lord, 1983). Supervisors observe a great deal of information about their subordinates. This information is encoded within evaluative "people" schemata, i.e. observations are used to form general, evaluative impressions of subordinates. This cognitive framework allows raters to make relatively easy comparative judgments among subordinates on general ability to perform their jobs. When required to make formal evaluations along specific performance dimensions, i.e. decode their impressions, supervisors first recall the most available information about their employee which is the general, comparative judgement. Then they access from "deeper" cognitive structures within this general impression, more specific dimension impressions. The resulting dimension evaluations would be made relative to the overall impression of the particular ratee, not in comparison to performance of other subordinates. In other words, general impressions are normative evaluations, while the specific impressions are ipsative

evaluations. Where specific incidents or behaviors are required, the rater can access still deeper, less available, information consistent with the dimension ratings.

Validity coefficients based on dimensional ratings would be expected to be lower than those based on an overall evaluation since the dimensional ratings are, in effect, made on different scales, each revolving around each rater's general impression of each ratee. Even where raters' overall impressions are accurate, for dimensional rating validities to approach those from overall ratings, raters would have to conceive the distributions of performance within dimensions similarly, and use those distributions consistently across different ratees, as well as have accurate representation of the performance of the employee encoded in the deeper dimension impression. Interestingly, our findings are consistent with Bingham's (1939) assessment of valid versus invalid halo. He found that raters were far better able to discriminate among ratees and showed a better agreement among themselves (inter-rater reliability) on complex qualities, including an overall rating on "personal fitness for the position," than on simpler, more observable qualities, such as appearance, voice, and command of language. Our findings would suggest that training programs focusing on teaching raters a common frame of reference for judging performance would be useful as a means of standardizing the cognitive structures across raters

(Bernardin and Buckley, 1981; McIntyre, Smith, and Hassell, 1984; Nathan and Alexander, 1985b). It would also suggest that evaluation techniques that require raters to agree on distributions of performance when rating different duties, such as Kane's (1984) Performance Distribution Assessment, would improve criterion measurement. Only in those duties where performance is easily observed and standards are generally obvious and agreed upon would we expect validity coefficients based on dimension (or behavior) ratings to be higher than those from the overall rating. Typing would appear to be one duty where this is true; so would duties with strong numerical ability requirements in which errors in calculations are easily detectable.

Since these findings make some very strong conclusions regarding the limited value of dimensional ratings as criteria for typical ability tests, replication is strongly advised. In addition, research needs to be conducted that tests the possibility that the individual dimensions may act as prompts to the rater, forcing him/her to consider all aspects of the rater's performance before making the overall rating. In this case, the process of actively working down through the dimensions may introduce a discipline which enhances the validity of the global rating.[1]

---

[1]The authors would like to thank Steve Wunder for suggesting this possibility.

In closing, we would like to add one other reason why we so strongly believe that general impressions and overall performance ratings are likely to be accurate. As Cooper (1981) has pointed out, employee selection and retention emphasize competency across all aspects of performance. If performance is unacceptable in some area, employees will typically receive training in that area or be terminated. Thus, halo can be caused not just by cognitive processes in the rater, but also by conditions in the organizational environment.

In light of these cognitive and environmental explanations, it is not surprising that the results of this study strongly suggest that ratings based on general impressions would seem to be the most appropriate (though still legally questionable) criteria for making the kinds of comparative judgements required in validation studies.

References

Bernardin, H. J., Buckley, M. R. (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.

Bingham, W. V. (1939). Halo invalid and valid. Journal of Applied Psychology, 23, 221-28.

Bittner, R. H., & Rundquist, E. A. (1950). The rank-comparison rating method. Journal of Applied Psychology, 34, 171-177.

Buckner, D. N. (1959). The predictability of ratings as a function of interrater agreement. Journal of Applied Psychology, 43, 60-64.

Cooper, W. H. (1981). Ubiquitous halo. Psychological Bulletin, 90, 218-244.

Feldman, J. M. (1986) A note on the statistical correction of halo error. Journal of Applied Psychology, 71, 173-176.

Harvey, R. J. (1982). The future of partial correlation as a means to reduce halo in performance ratings. Journal of Applied Psychology, 67, 171-176.

Henik, A. and Tzelgov, J. (1985). Control of halo error: A multiple regression approach. Journal of Applied Psychology, 70, 577-580.

Hulin, C. L. (1982). Some reflections on general performance dimensions and halo rating error. Journal of Applied Psychology, 67, 165-170.

Kane, J. S. (1984). Performance distribution assessment: A new breed of appraisal methodology. In Bernardin, H. J., and Beatty, R. W:,

Performance Appraisal: Assessing Behavior at Work. Boston: Kent Publishing.

Lance, C. E., & Woehr, D. J. (1986). Statistical control of halo: A clarification from two cognitive models of the performance appraisal process. Journal of Applied Psychology, 71, 679-685.

Landy, F. J., Vance, R. J., & Barnes-Farrell, J. L. (1982). Statistical control of halo: A response. Journal of Applied Psychology, 67, 177-180.

Landy, F. J., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W. (1980). Statistical control of halo error in performance ratings. Journal of Applied Psychology, 65, 501-506.

McIntyre, R. M., Smith, D. E., and Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Mossholder, K. W., & Giles, W. F. (1983). The use of partial correlation to control halo in performance ratings. Educational and Psychological Measurement, 43, 977-984.

Murphy, K. R. (1982) Difficulties in the statistical control of halo. Journal of Applied Psychology, 67, 161-164.

Nathan, B. R., & Alexander, R. A. (1985). The predictability and substitutability of criteria for clerical workers: A meta-analytic investigation. Proceedings of the 28th Annual Conference of the Midwest Academy of Management. Champaign, IL. (a)

Nathan, B. R., and Alexander, R. A. (1985). The role of inferential accuracy in performance rating. Academy of Management Review, 10, 109-116. (b)

Nathan, B. R. & Cascio, W. F. (1986). Legal and technical standards for performance assessment. In Berk, R. A. (ed.) Performance assessment: Methods and Applications. Baltimore, MD: Johns Hopkins University Press.

Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68, 102-114.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Rush, C. H., Jr. (1953). A factorial study of sales criteria. Personnel Psychology, 6, 9-24.

Schmidt, F. L., Hunter, J. E., and Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 66, 166-185.

Table 1

Correlations and Partial Correlations Among Dimension Ratings

	Dimensions[a]																	
-	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	-	.43	.30	.48	.45	.54	.51	.38	.65	.39	.36	.48	.41	.51	.52	.33[b]	.57[b]	.64
2	-.04	-	.44	.40[b]	.56	.54	.49	.59	.52	.26	.49	.59	.62	.49	.47	.26[d]	.44[b]	.69
3	-.20*	.12	-	.29[b]	.57	.18	.48	.49	.40	.23[c]	.49	.43	.51	.26[c]	.49	.10[e]	.28[d]	.58
4	.21[+]	.00	-.25*	-	.59	.63	.50	.48	.57	.62	.46	.28	.20	.41	.48	.46	.33[b]	.66
5	.02	.08	-.23*	.20*	-	.59	.57	.58	.62	.50	.56	.49	.45	.47	.49	.43	.49	.73
6	.05	.01	.04	.27***	.04	-	.67	.62	.61	.51	.52	.47	.46	.54	.69	.42	.57	.78
7	.08	-.02	-.02	.13	.17*	.25***	-	.69	.63	.43	.54	.50	.49	.55	.54	.56	.49	.73
8	-.16*	.20*	.13	.11	.16*	.16*	.39	-	.59	.41	.52	.55	.59	.55	.55	.41	.38	.70
9	.29***	-.06	-.09	.16[+]	.17*	.08	.20**	.12[+]	-	.51	.52	.55[b]	.50[b]	.49	.58	.37	.53	.75
10	.05	-.02	-.21[+]	.42***	.23**	.13	.06	.06	.15[+]	-	.41	.24[b]	.34[b]	.41	.44	.50	.56	.52
11	.00	.14[+]	.26**	.07	.22**	.10	.24***	.12[+]	.11[+]	.18*	-	.58	.45	.48	.51	.37	.42	.65
12	.12	.28***	.16	-.06[+]	.13	.01	.21**	.21*	.17	.01	.32***	-	.46	.48	.43	.14[e]	.51	.62
13	-.09	.33***	.24*	-.20[+]	.02	-.03	.06	.30***	.11	.17[+]	.19*	.17*	-	.43	.49	.33[b]	.24[e]	.60
14	.11	.09	-.11	.04	.06	.15	.29**	.32***	.12	.19[+]	.20*	.21*	.16	-	.42	.47	.61	.62
15	-.07	-.03	.13	-.01	.04	.30**	-.06	.06	.01	.07	.14*	.02	.06	.11	-	.36	.40[b]	.74
16	.06	-.06[+]	-.20	.16	.10	.04	.35[+]	.13	-.04	.25*	.08	-.16	.08	.22*	-.06	-	.51[b]	.54
17	.29	-.23[+]	-.12	.00	.01	.16	.24[+]	-.15	.14	.35**	.01	.15	-.06	.38**	.18	.32[+]	-	.62

Note: N-size varies from 32 to 230. Observed correlations are above the diagonal, all are significant at $p < .001$ except those indicated by 'b' ($p < .01$), 'c' ($p < .05$), 'd' ($p < .10$), 'e' (ns). Partial correlations are below the diagonal, significance is indicated in traditional manner: '+' $p < .10$, '*' $p < .05$, '**' $p < .01$, '***' $p < .001$.

[a]Dimensions: 1 = Typing, 2 = Performing Receptionist Activities, 3 = Ordering Supplies and Monitoring Inventory, 4 = Analyzing Data, 5 = Copying and Recording Data, 6 = Solving Job-related Problems, 7 = Responding to Written Requests, 8 = Responding to Oral Requests, 9 = Checking and Reviewing Written Information, 10 = Performing Arithmetical Calculations, 11 = Classifying and Filing Information, 12 = Distributing Material, 13 = Scheduling Events and Activities, 14 = Operating Office Machines, 15 = Establishing Priorities and Scheduling Own Work, 16 = Preparing Written Reports, 17 = Keypunching/Data Entry, 18 = Overall Job Performance.

Author Notes

The authors would like to thank James Dahl of the University of Missouri-St. Louis Department of Psychology, for assisting in the analysis of this data.

Correspondences should be addressed to Barry R. Nathan, Department of Management and Organization, School of Business Administration, University of Southern California, Los Angeles, CA 90089-1421.

Parts of this study have been submitted to the 1987 National Meeting of the American Psychological Assoication for review and presentation.

Table 2

Validation Results for Predictors Correlated with Observed Ratings and Ratings "Statistically Controlled for Halo"

		Predictors							
		Clerical Ability		Numerical Ability		Verbal Ability		General Mental Ability	
Duty	(N)	Obs.[a]	Res.[b]	Obs.	Res.	Obs.	Res.	Obs.	Res.
Overall Job Performance	(279)	.24***	--	.18*	--	.25***	--	.27***	--
Typing	(176)	.18*	.04	.05	-.08	.36***	.26**	.26**	.11
Receptionist Activities	(177)	.18*	.03	.11	-.02	.23**	.09	.13	-.07
Ordering Supplies	(168)	.13	-.01	.18*	.09	.15	.00	.09	-.08
Analyzing Data	(102)	.27***	.16*	.25***	.18*	.16*	-.01	.29***	.15*
Copying/Recording Data	(222)	.25***	.11	.21**	.11	.19**	.02	.30***	.14*
Solving Job Related Problems	(232)	.22***	.05	.21**	.10	.27***	.12	.30***	.13*
Responding to Written Requests	(193)	.14*	-.04	.16*	.05	.21**	.04	.24***	.06
Responding to Oral Requests	(237)	.26***	.14	.22***	.13*	.25***	.10	.28***	.12*
Checking Written Information	(239)	.16**	-.02	.22***	.13*	.20**	.02	.30***	.15*
Performing Arithmetic Calculation	(148)	.17*	.05	.28***	.22**	.14	.01	.32***	.20**
Classifying/Filing	(252)	.21***	.08	.23***	.15*	.21***	.06	.22***	.06
Distributing Material	(236)	.11	-.05	.13*	.02	.23***	.10	.12	-.06
Scheduling Events	(147)	.17*	.03	.10	-.01	.18*	.03	.10	-.08
Operating Office Machinery	(149)	.13	-.03	.22**	.14	.23**	.09	.29**	.15*
Establishing Priorities	(250)	.19*	.02	.09	-.07	.12	-.10	.14*	-.09
Preparing Written Reports	(96)	.15	.03	.19	.11	.11	-.03	.20	.06
Keypunching/Data Entry	(93)	.09	-.08	.03	-.10	.12	-.05	.09	-.10

Note: Significance varies depending on N-size and round-off error. '*' $p < .05$, '**' $p < .01$, '***' $p < .001$

[a] Correlations based on observed ratings scores.

[b] Part correlations based on residual rating scores where overall rating has been statistically removed in order to "statistically control for halo."

Table 3

Incremental Variance from Hierarchical Regressions of Predictor Scores on Job Duty and Overall Performance Ratings

Duty		Predictors							
		Clerical Ability		Numerical Ability		Verbal Ability		General Mental Ability	
	(N)	Duty[a]	Overall[b]	Duty	Overall	Duty	Overall	Duty	Overall
Typing	(176)	.00	.02*	.01	.04*	.07***	.00	.01	.02
Receptionist Activities	(177)	.00	.03*	.00	.02*	.01	.02	.00	.06***
Ordering Supplies	(168)	.00	.04*	.01	.01	.00	.04*	.01	.07***
Analyzing Data	(162)	.02*	.01	.03*	.00	.00	.04*	.02*	.01
Copying/Recording Data	(222)	.01	.01	.01	.00	.00	.03*	.02*	.01
Saving Job-related Problems	(232)	.00	.01	.01	.00	.01	.00	.02*	.00
Responding to Written Requests	(193)	.00	.04**	.01	.00.	.00	.02	.00	.02*
Responding to Oral Requests	(237)	.02*	.01	.02*	.00	.01	.01	.02*	.01
Checking Written Information	(239)	.03**	.00	.02*	.00	.00	.02	.02*	.01
Performing Arithmetic Calculations	(148)	.00	.03*	.05**	.00	.00	.04**	.02	.04**
Classifying/Filing	(252)	.01	.02*	.02*	.00	.00	.02*	.00	.03**
Distributing Material	(236)	.00	.05***	.00	.02*	.01	.02*	.00	.06***
Scheduling Events and Activities	(147)	.00	.03*	.00	.02	.00	.03*	.01	.07***
Operating Office Machinery	(147)	.00	.04*	.02	.00	.01	.02	.02	.01
Establishing Priorities and Scheduling Own Work	(252)	.00	.02*	.00	.03**	.01	.06***	.01	.06***
Preparing Written Reports	(96)	.00	.03	.01	.01	.00	.05*	.00	.04*
Keypunch/Data Entry	(93)	.01	.05*	.01	.04*	.00	.05*	.00	.08**

Note: Significance varies depending on N-size and round-off error. '*' $p < .05$, '**' $p < .01$, '***' $p < .001$

[a] Incremental R^2 due to specific duty rating beyond that due to overall performance rating.

[b] Incremental R^2 due to overall duty rating beyond that due to specific duty rating.