Journal of Applied Psychology
1990, Vol. 75, No. 1, 28-42

0021-9010/90/$00.75

Individual Differences in Output Variability as a Function of Job Complexity

John E. Hunter
Michigan State University

Frank L. Schmidt and Michael K. Judiesch
Department of Management and Organizations
University of Iowa

The hypothesis was tested that the standard deviation of employee output as a percentage of mean output (SD_p) increases as a function of the complexity level of the job. The data examined were adjusted for the inflationary effects of measurement error and the deflationary effects of range restriction on observed SD_p figures, refinements absent from previous studies. Results indicate that SD_p increases as the information-processing demands (complexity) of the job increase; the observed progression was approximately 19%, 32%, and 48%, from low to medium to high complexity nonsales jobs, respectively. SD_p values for sales jobs are considerably larger. These findings have important implications for the output increases that can be produced through improved selection. They may also contribute to the development of a theory of work performance. In addition, there may be implications in labor economics.

One of the factors determining the utility or economic value of personnel selection is the variability of output of employees selected randomly from the applicant pool. This variability has typically been estimated as the standard deviation of the dollar value of output, symbolized as SD_y (Schmidt, Hunter, McKenzie, & Muldrow, 1979). However, it can also be expressed as the ratio of the standard deviation of output to mean output, symbolized as SD_p (Schmidt & Hunter, 1983). Use of SD_y leads to utility expressed in dollars, whereas use of SD_p leads to utility expressed as the percentage increase in output (Hunter & Schmidt, 1983; Schmidt, Hunter, Outerbridge, & Trattner, 1986; Schmidt, Mack, & Hunter, 1984). In addition to this practical value, knowledge of the extent and magnitude of individual differences in output (often referred to as "productivity" but actually production) is of general theoretical interest to the field of applied differential psychology. Job performance or output is probably the most important dependent variable in industrial/organizational psychology. The question of how much employees in the same job typically differ in output is central to an individual-differences approach to job performance.

How much do workers differ in output? The answer often desired has a form such as "Top workers are 3 times more productive than bottom workers," or "Top workers are 50% more productive than average workers." Such statements require that performance be measured on a ratio scale, to allow the computation of the ratio of the performance of higher output workers to lower output workers. Extremely rare workers can always be found farther and farther out on either end of the performance continuum as time goes on. For a normal distribution, there is no population maximum or minimum performance. Thus, one must instead create an essential maximum by choosing a top percentage—the top 1%, for example. Similarly, one can create an essential minimum by choosing a bottom percentage—the bottom 1%, for example. For a normal distribution, top and bottom performance can be computed given the mean and standard deviation. For example, the top 1% averages about 2.67 standard deviations above the mean, whereas the bottom 1% averages about 2.67 standard deviations below the mean. (The score that cuts off the extreme 1% is 2.33 *SD*s from the mean; but the *average z* score for the top 1% is larger. This average is ϕ/p, where ϕ = the ordinate in N(0, 1) at the point of cut, and p is the proportion in the extreme group, here .01.)

For a given job, the mean and standard deviation can be on any scale as long as it is a ratio scale. Scale units can vary from "dresses sewn" on one job to "cars repaired" on another. However, a common metric is necessary to make figures comparable across jobs. One such metric is provided by expressing the standard deviation of performance as a percentage of mean performance, that is, by using the coefficient of variation times 100. This scale is obtained by multiplying each output level by the constant $100/M$, where M is mean output for the job in question. This transformation rescales the output unit so that mean output is 100 while preserving the ratio property of the scale. For example, suppose that mean daily output were 50 items produced and that the top and bottom 1% of workers produced 75 and 25 items, respectively, yielding an extreme ratio of 75/25 = 3, or 3 to 1. The percentage performance scores would be 150 and 50, maintaining the ratio of 150/50 = 3, or 3 to 1. The standard deviation of the common scale is the ratio of the original metric standard deviation to the original metric mean, times 100. This output standard deviation ratio is symbolized as SD_p.

Schmidt and Hunter (1983) reviewed the literature for studies that reported SD_p or data from which SD_p could be calculated. They found 40 data sets: 29 studies of production in ordinary pay conditions and 11 studies of piece-rate systems. For ordinary (non-piece-rate) pay conditions, their data suggested a "conservative" baseline of 20% for the standard deviation of output. The findings in that study were based mostly on blue-collar skilled and semiskilled workers and routine clerical jobs.

Correspondence concerning this article should be addressed to Frank L. Schmidt, Department of Management and Organizations, College of Business Administration, University of Iowa, Iowa City, Iowa 52242.

Schmidt and Hunter suggested that the relative variation might be higher for more complex jobs. The present study reports new data for more complex jobs and tests that hypothesis. Because over 95% of the jobs in the economy are non-piece-rate jobs (i.e., have nonincentive-based compensation systems; Bureau of National Affairs, Inc., 1983), the primary focus of our study was on those jobs. However, many higher level jobs (e.g., physician, dentist, attorney) and sales jobs (e.g., life insurance sales) contain linkages between output and earnings that are essentially inherent in the occupations. Such jobs were included in this study.

Our study differs from that of Schmidt and Hunter (1983) in the following two additional respects: (a) Formal corrections for range restriction in the incumbent SD_p values are made, and (b) adjustments are made to correct for the effects of measurement error in the output measures.

Range Restriction: The Applicant Standard Deviation

In evaluating the utility of programs applied to incumbent employees (e.g., training or performance evaluation programs), the reference population is incumbent workers, and the relevant output standard deviation ratio (SD_p) is that for incumbent workers (Hunter & Schmidt, 1983; Schmidt, Hunter, & Pearlman, 1982). However, in personnel selection, the reference population is the applicant population rather than the incumbent workers. In particular, utility analysis for personnel selection requires the applicant population output standard deviation ratio (Schmidt et al., 1979). Empirical studies are of necessity conducted on incumbent workers. Thus, observed SD_p values are those for incumbent workers rather than for applicants. The standard deviation for incumbent workers is subject to restriction in range caused by selective hiring, selective promotion of better workers, and selective termination of poorer workers. At present, quantitative data are available that allow correction for the average level of range restriction caused by the selection of workers on general cognitive ability. These corrected estimates systematically underestimate the actual applicant output standard deviation ratio because the standard deviation is not corrected for restriction caused by selective promotion and termination.

Schmidt and Hunter (1983) reported observed SD_p values. They noted that their incumbent standard deviations were restricted but presented no quantitative estimate of the effect of restriction. In our study, a method is presented for correcting incumbent standard deviations to estimate applicant standard deviations. As would be expected, applicant standard deviations are found to be larger than incumbent standard deviations.

Effects of Unreliability on SD_p Values

Unreliability in measures of job performance reduces observed validities from their true value, and correcting for measurement error in the criterion yields an estimate of true validity that is larger than the observed (i.e., initially computed) validity. However, in the case of SD_p, measurement error has the opposite effect: It inflates rather than attenuates the estimate of SD_p. SD_p is the ratio of the standard deviation of output to mean output, that is, SD/M. Measurement error creates no bias in the denominator, M. However, the numerator is biased upward. The observed variance of the output measure is the sum of the true variance of output plus the variance of measurement errors, that is, $S_T^2 + S_e^2$. The true variance is smaller, i.e., S_T^2. Thus, correcting for unreliability *reduces* the estimate of SD_p. For example, if an SD_p estimate of .30 is based on 1 week of measured output on the job, and if the correlation between any 2 weeks of output is .75 (reflecting intraindividual output variability), then the corrected estimate of SD_p is $(.75)^{1/2}$ (.30), or .26. That is, unreliability of output measures causes the observed SD_p to be 15% too large, and correction eliminates this bias. Our earlier study (Schmidt & Hunter, 1983) overlooked this fact.

Method

The literature, both in industrial/organizational psychology and in other areas, was extensively searched for new studies containing information allowing computation of SD_p. This literature consists of (a) studies that report the mean and standard deviation of actual employee production or output and (b) studies that report the ratio of actual output of highest producing employees to actual output of lowest producing employees. In the latter type of study, given the assumption of normality of the full output distribution, the standard deviation as a percentage of mean output can be computed for all studies that give the total sample size by using the formula given in Schmidt and Hunter (1983, p. 408). However, for reasons described later, the normality assumption is not plausible and was not made for high-complexity jobs or for sales jobs. This constraint eliminated one study of a high-complexity job: the Rimland and Larson (1986) study of computer programmers. (In this study, the output of the highest producing programmers averaged 16 times that of the lowest producers.) All other studies located beyond those included in Schmidt and Hunter (1983; i.e., "new studies"), regardless of level of job studied, reported means and standard deviations, allowing direct computation of the ratio. Schmidt and Hunter (1983) used formula computation of SD_p only for low- or medium-complexity jobs; at these levels, the normality assumption is quite plausible, as discussed later.

In addition to studies reporting on-the-job output, studies using work sample measures based on ratio scales of output could be used. Job sample measures were considered not to have ratio scale properties if the scoring was based on *ratings* of quality or quantity of output. When the score was based on a count of output (either total or acceptable output), the scale was considered to have ratio properties and was included. Such studies provided data on a number of medium-complexity jobs and one high-complexity job. Finally, in some high-complexity jobs, earnings are very closely tied to output and therefore the SD_p value can be based on means and standard deviations of earnings in the occupation. Surveys reporting such data are available for various professional occupations. This approach allowed computation of income-based SD_p values for attorneys, physicians, and dentists. In the case of other professions reporting earnings distributions (e.g., accounting), compensation appeared to be based on salary and therefore earnings were not directly and immediately dependent on output; therefore, such data were excluded from the study.

Owens (1987) reported a national survey of the earnings of nonsalaried physicians. The figures used in the ratio *SD/M* for our study are those across regions and specialties. This figure is 52.8. When medical specialty is held constant, SD_p varies from 41.3 to 56.7 across the 13 specialties, with a mean of 49.6. Regional differences in earnings are small; median values for the 9 regions vary from $97,500 to $125,630, with a mean of $111,447. Theodore and Sutter (1967) reported statistics on number of patient visits per physician, also based on national data. The data we used from this study were those for physicians who re-

ported 40 or more hours of patient care per week; the resulting SD_p was 63.7. For all physicians, this value was 75.0. The SD_p figures ranged from 65.3 to 77.4 across geographic regions. Across specialties, SD_p ranged from 57.6 for general practitioners to 79.1 for surgeons, with a mean of 66.6.

The American Dental Association (1969a) reported a national survey of the earnings of nonsalaried dentists. The observed SD_p figure for full-time dentists nationally is 51.4. These data are not presented separately by geographic region or specialty. A separate article (American Dental Association, 1969b) reported data on number of patient visits per year for the same large national sample (N = 4,023). The observed SD_p for these figures is 43.1.

Altman & Weil, Inc. (1981) reported national data on the earnings of attorneys. Nonpartners in law firms are frequently on salary, which may not immediately and directly reflect output. We used only the data for partners and shareholders, which produced an observed SD_p of 50.3 (N = 4,659). The figure for only those attorneys with 25–30 years of experience (the group with the highest earnings) was very similar: 51.4 (N = 409). No data comparable with number of patient visits for physicians and dentists could be found for attorneys (see Appendix A).

For all three of these professions, the surveys included a national cross-section of the occupation that cut across employers. Thus, survey samples are representative of the potential applicant pool for these occupations. That is, the resulting SD_p values apply to the field as a whole (i.e., the potential applicant pool) rather than being an average of within-employer values (i.e., incumbent values). In view of this, corrections for range restriction were not appropriate for these three professions and were not applied.

In all the studies reviewed, employee output was self-paced; in none of the studies did employee rate of production appear to be constrained by the production technology, as it would be in the case of an assembly line. A number of studies presented findings separately for experienced employees and all employees; in such cases, only the results for the experienced employees were used.

Jobs were assigned to low-, medium-, and high-complexity levels on the basis of a modification of Hunter's (1980) system. That system is based on the *Dictionary of Occupational Titles's* (U.S. Department of Labor, 1977) data–people–things job analysis procedure and consists of five complexity levels.[1] The first two levels (managerial/professional and complex technical set-up work) correspond to our high-complexity level. His third level, which includes most skilled crafts, technician jobs, first-line supervisors, and lower level administrative jobs, corresponds to our medium-complexity level. Finally, Hunter's (1980) two lowest complexity jobs (essentially semiskilled and unskilled) correspond to our low-complexity level. For purposes of this study, sales jobs were kept in a separate category and were not assigned to complexity levels.

It quickly became apparent that many studies with otherwise usable data did not report reliabilities for output measures. It was also clear that reliabilities could vary by type of output measure: (a) counts of output on the job, (b) job sample measures, and (c) measures of sales. Finally, as would be expected, reliabilities varied with the amount of time over which output was measured. For example, the correlation between two 4-week periods was larger than the correlation between two 1-week periods. We therefore recorded reliabilities separately for the three types of measures. Within each type of measure (e.g., sales) we used the Spearman-Brown formula to adjust the reliabilities up or down to correspond to a constant time period (1 week or 4 weeks). (This step was not necessary for job sample measures.) The mean value for this constant time period was computed and this mean was the basis for the correction applied, adjusting for the time period of measurement. For example, the mean reliability for a 1-week measure of output on the job was .55. (That is, the average correlation between any 2 weeks was .55.) If a study reported output measured over a 5-week period, this figure, adjusted by the Spearman-Brown formula to 5 weeks (.86) was used to make the correction if the study reported no reliability. If the study reported a reliability estimate, the estimate from that study was used. There is some evidence that reliabilities of output measures are higher for piece-rate than for non-piece-rate jobs (Rothe, 1978). Therefore, reliabilities from piece-rate jobs were not tabulated.

The method used to estimate (unrestricted) applicant SD_p values is explained in Appendix B. This method takes advantage of the fact that accurate empirical estimates of the average range restriction on general mental ability are available. This information is used in conjunction with the best available meta-analytic estimate of the mean correlation between general mental ability and job performance (measured by using content-valid job sample tests) to determine the reduction in the incumbent SD_p due to selection on general mental ability. The resulting adjustment corrects only for the effects of selective hiring; it does not adjust for the effects of selective promotion of higher performing employees or for the selective voluntary and involuntary turnover of poorer performing employees. Thus, the correction is an undercorrection.

Results and Discussion

Reliability of Output Measures

Tables 1, 2, and 3 show the results of the reliability analyses. In Table 1, it can be seen that 12 studies reported reliabilities of counts of output on non-piece-rate jobs. Reliabilities were reported for time periods ranging from 1 week to 26 weeks. The average reliability for 1 week was .55; this figure was used with the Spearman-Brown formula to compute expected average reliabilities for each of the time periods shown. It is clear from Table 1 that the reliability of actual counts of employee output (often referred to as *objective criteria*) over periods of a week or so is less than is often implicitly assumed. The 1-week figure of .55 is similar to the average interrater reliability of ratings based on two raters (King, Hunter, & Schmidt, 1980).

Table 2 shows that 10 estimates of the reliability of job sample tests were found and that the average reliability was .78. The duration of job samples was typically under one day and often only a few hours. Thus, on a per-unit time basis, carefully constructed and administered job samples appear to be more reliable than counts of actual output taken on the job. However, the job sample reliabilities were typically computed at one point in time and thus do not reflect any instability in performance over time (i.e., transient errors) that may exist. Meaningful comparisons with the reliabilities of output measures must await the availability of test–retest reliability estimates for job sample measures.

Table 3 shows the reliability findings separately for life insurance sales and other sales. In both cases, reliabilities are lower for a 4-week period for sales than for a 1-week period for non-sales output (Table 1). The average for life insurance sales for 4

[1] In Hunter's (1980) system, the two highest complexity levels consist of (a) jobs with a code of 0 or 1 on the Data dimension (e.g., scientists, executives) and (b) jobs with a code of 0 on the Things dimension (e.g., computer trouble shooters). (These two categories are considered essentially equal in complexity.) The third level of complexity consists of jobs with codes on the Data dimension of 2, 3, or 4 (e.g., welder, auto mechanic, general clerk). The next level of complexity is made up of jobs with data codes of 5 or 6 (e.g., truck driver, assembler, file clerk). Finally, the lowest level of complexity is represented by jobs with a code of 6 on the Things dimension. On all dimensions, higher codes indicate lower levels (U.S. Department of Labor, 1977). The People dimension is not used because codes on that dimension were found not to moderate General Aptitude Test Battery validities.

Table 1
Reliability of Output Measures: Nonincentive Systems

Study	Job	Time period[a]	No. rs	N	Reliability[b] Time period (weeks) 1	2	3	4	5	8	13	26	52
Gaylord, Russell, Johnson, & Severin (1951)	Clerks	days	Many	59	.66							.98[d]	
Rothe (1947)	Machine operators	2 weeks	3	130	.49	.66[d]							
Rothe & Nye (1958)	Coil winders	1 week	37	27	.60[d]								
Rothe & Nye (1961)	Machine operators	1 week	10	37	.48[d]								
Rothe & Nye (1961)	Machine operators	1 week	11	61	.53[d]								
Rothe (1970)	Welders	1 week	47	25	.52[d]								
Rothe (1978)[c]	Butter wrappers	2 weeks	2	8	.46	.63[d]							
Tiffin & McCormick (1965)	Electrical workers	5 weeks	1	79	.47				.87[d]				
Hay (1943)	Machine bookkeeper	4 days	1	39	.82								
Ledvinka, Simonet, Neiner, & Kruse (1983)	Claims evaluator	1 month	1	15	.85					.98[d]			
Hearnshaw (1937)	Paper sorters	3 months	1	18	.18						.74[d]		
Validity Information Exchange No. 11-27 (1958)	Grid operator	2 weeks	1	63	.47	.64[d]							
Sample size weighted mean R_{YY}					.55	.71	.79	.83	.86	.91	.94	.97	.99

[a] This was the time period used in the study to compute the initial correlation. [b] Reliability was based on average correlations when more than one correlation between time periods was given; values that were based on more than 1 week were adjusted to one week by using the Spearman-Brown formula. [c] These data were said by the author to be based on the data used in Rothe (1946). [d] This was the value reported in the study; other values in column 1 were calculated by using the (reversed) Spearman-Brown formula.

weeks is only .23, and for other sales it is .39. These figures are potentially deceptive, however, in that in all studies located, sales measures were taken over longer time periods, ranging from 13 weeks to 52 weeks. Thus, the reliabilities of the sales measures actually used in these studies (and in our study) were considerably higher.

Reanalysis of Schmidt and Hunter (1983) Data

Schmidt and Hunter (1983) did not segregate jobs by complexity level nor did they separate out sales jobs. They obtained an average SD_p of 20.0% for non-piece-rate jobs. Table 4 shows a reanalysis of their 29 SD_p estimates for non-piece-rate jobs broken down in this manner and corrected for the inflationary effects of unreliability. Notes to Table 4 explain the details of the reliability corrections. Fifteen of their reports were for routine blue-collar jobs. For these jobs, the average observed SD_p was 18.5%, close to their overall 20% mean. But this figure shrinks to 14.1% after correction for unreliability. (These figures have not yet been corrected for range restriction.) A similar pattern holds for the seven estimates from routine clerical jobs. These two classes, taken together, make up the low-complexity cate-

Table 2
Reliability of Output Measures: Job Sample Studies

Study	Job	N	Reliability
Stead & Shartle (1940)	Typists	222	.96
Whipple, Baldin, Mager, & Vineberg (1969)			
Group 1	Radar mechanics	107	.71
Group 2	Radar mechanics	51	.71
U.S. Postal Service (1981)	Mail handler (sorting)	373	.88
U.S. Postal Service (1981)	Mail handler (moving)	373	.64
U.S. Postal Service (1981)	Mail carrier	374	.94
U.S. Postal Service (1981)	Mail distribution clerk	417	.96
Campbell, Crooks, Mahoney, & Rock (1973)	Cartographer	443	.49
Corts, Muldrow, & Outerbridge (1977)	Customs inspector	186	.80
Trattner, Corts, van Rijn, & Outerbridge (1977)	Claims authorizer	233	.72
Sample size weighted mean R_{YY}			.78

Table 3
Reliability of Output Measures: Sales Jobs

Study	Job	Time period[a]	No. rs	N	Reliability[b] Time period (weeks) 4	8	13	26	41	52
			Life insurance sales							
Manson (1925)	Life insurance agent	1 year	1	1,528	.15					.69[c]
Strong (1935)	Life insurance agent	1 year	4	102	.21					.78[c]
Kahn & Hadley (1949)	Life insurance agent	13 weeks	1	65	.27		.55[c]			.83[c]
Brown (1981)										
Company 1	Life insurance agent	6 months	1	3,590	.26					.82[c]
Company 2	Life insurance agent	6 months	1	768	.26					.82[c]
Company 3	Life insurance agent	6 months	1	949	.24					.80[c]
Company 4	Life insurance agent	6 months	1	1,606	.25					.81[c]
Company 5	Life insurance agent	6 months	1	752	.25					.81[c]
Company 6	Life insurance agent	6 months	1	893	.25					.81[c]
Company 7	Life insurance agent	6 months	1	606	.25					.81[c]
Company 8	Life insurance agent	6 months	1	793	.16					.72[c]
Company 9	Life insurance agent	6 months	1	771	.16					.72[c]
Company 10	Life insurance agent	6 months	1	658	.18					.74[c]
Company 11	Life insurance agent	6 months	1	661	.21					.78[c]
Company 12	Life insurance agent	6 months	1	406	.26					.82[c]
Sample size weighted mean R_{YY}					.23	.37	.49	.66	.75	.80
			Other sales							
Rush (1953)	Office machinery sales	1 month	1	100	.08				.47[c]	
Weekley & Gier (1987)	Department store sales clerks	6 months	1	573	.44					.91[c]
Sample size weighted mean R_{YY}					.39	.56	.68	.81	.87	.89

[a] The time period used to compute the initial correlations. [b] Based on average correlations when more than one correlation between time periods is given; reported values based on more than 4 weeks are adjusted to four weeks using the Spearman-Brown formula. [c] Value reported in study; other values were calculated using the (reversed) Spearman-Brown formula.

gory. As expected, medium-complexity jobs show both higher observed and adjusted SD_p values. The two sales categories have even larger SD_p values. Thus, there is evidence in the original Schmidt and Hunter (1983) data that SD_p values vary with job complexity and with the sales–nonsales distinction. There is also evidence that measurement error in output counts inflates observed SD_p values.

SD_p *Values From New Studies*

In this section we present findings from studies located subsequent to Schmidt and Hunter (1983). Table 5 shows the findings for newly located nonsales studies, presented in the same way as in Table 4. The same pattern noted in Table 4 is observed in this independent set of studies. SD_p mean values, both observed and corrected for measurement error, are larger for medium-complexity jobs than for lower-complexity jobs. The corrected mean values are almost identical in the two tables for the two classes of low-complexity jobs: routine blue-collar and routine clerical. The mean values for the two classes of medium-complexity jobs, crafts and clerical with decision making, are slightly larger in Table 5 than in Table 4. Unlike Table 4, Table 5 presents findings for high-complexity (professional) jobs. As predicted, the mean SD_p value is largest of all for this complexity level: The average corrected value is 46.2%. Table 6 shows the findings for new studies of sales jobs; results are presented separately for insurance sales and other sales. Even after correcting for unreliability, the mean SD_p value for life insurance sales is still very large: 96.6%. The average variability of output across incumbents is unusually large in this occupation. However, even for noninsurance sales, the average value (42.3%) is fairly large, being only a little less than that for high-complexity (professional) jobs in Table 5 (46.2%). The average for noninsurance sales is based on five different kinds of sales jobs. One of these jobs (sales account manager) has a SD_p (76.3) much closer to the life insurance sales mean (96.6) than the others. The other four are more homogeneous and are similar to the two SD_p values for sales jobs in Table 4. This suggests that other non–life insurance sales jobs might exist that have SD_p values closer to the average for the life insurance sales job. Future studies should test this hypothesis.

Combined Findings

Table 7 combines the figures in Tables 4, 5, and 6 and corrects for range restriction to estimate applicant pool SD_p values. The

Table 4
Incumbent Output Standard Deviations as Percentage of Mean Output (SD_p): By Job Complexity and Groups (Adapted from Schmidt & Hunter, 1983)

Study	Occupation	*N*	Observed incumbent SD_p	Time period (weeks)	Reliability	Reliability corrected SD_p
		Low complexity				
Routine blue collar						
Rothe (1946)[a]	Dairy workers	8	23.2	2	.63	18.4
Rothe (1947)[b,c]	Machine operators	130	25.2	2	.66	20.5
Rothe & Nye (1958)[b]	Industrial workers	27	19.4	1	.60	15.0
Rothe & Nye (1961)						
1958[b]	Machine operators	37	16.9	1	.48	11.7
1960[b]	Machine operators	61	10.3	1	.53	7.5
Tiffin (1947)[d]	Electrical workers	33	17.8	NR	.55	13.2
Barnes (1958)[e]	Assembly workers	294	14.0	4	.83	12.8
Stead & Shartle (1940)[d]	Lamp shade manufacture	19	11.6	NR	.55	8.6
Lawshe (1948)[d]	Wool pullers	13	20.3	NR	.55	15.1
Wechsler (1952)						
Group 1[e]	Machine sewing	101	19.7	1	.55	14.6
Group 2[d]	Electrical workers	100	12.9	NR	.55	9.6
Group 3[d]	Electrical workers	65	17.1	NR	.55	12.7
McCormick & Tiffin (1974)						
Group 1[d]	Cable workers	40	23.8	NR	.55	17.7
Group 2[d]	Electrical workers	138	9.0	NR	.55	14.1
Group 3[d]	Assemblers	35	26.4	NR	.55	19.6
No. studies and average		15	18.5			14.1
Routine clerical						
Klemmer & Lockhead (1962)[e]	Card punch operators	NR	11.6	52	.99	11.5
Klemmer & Lockhead (1962)[e]	Proof machine operators	NR	13.5	52	.99	13.4
Stead & Shartle (1940)[f]	Typists	616	18.7	JS	.99	18.6
Group 2	Card punch operators					
	Day shift[d]	113	14.4	NR	.55	10.7
	Night shift[d]	121	17.4	NR	.55	12.9
Group 4[d]	Card punch operators	62	29.1	NR	.55	21.6
Lawshe (1948)[g]	Cashiers	29	19.6	JS	.78	17.3
No. studies and average		7	17.8			15.1
		Medium complexity				
Crafts						
Rothe (1970)[b]	Welders	25	19.0	1	.52	13.7
Evans (1940)[d]	Handcrafters	NR	23.0	NR	.55	17.1
Lawshe (1948)[h]	Drilling	11	33.0	6	.88	31.0
No. studies and average		3	25.3			20.6
		Sales				
Stead & Shartle (1940)[i]	Sales clerks	153	33.5	NR	.44	22.2
Lawshe (1948)[i]	Sales clerks	18	54.2	4	.44	36.0
No. studies and average		2	43.9			29.1

Note. NR = not reported; JS = job sample.
[a] R_{YY} for these data is given in Roth (1978); see Table 1. [b] R_{YY} was given in the study for the correct time period. [c] This was the average of values for the same subjects for three time periods, reported separately in Schmidt and Hunter (1983). [d] R_{YY} was not given; the 1-week average from Table 1 was used. [e] R_{YY} was not given; the mean from Table 1 for that time period was used. [f] R_{YY} was from Stead and Shartle (1940). [g] R_{YY} was not given; the mean from Table 2 was used. [h] R_{YY} was not given; it was computed from the Table 1 mean for 1 week, using the Spearman-Brown formula. [i] R_{YY} was not given; R_{YY} for 4 weeks from Weekley and Gier (1987) was used.

mean applicant SD_p values are similar for the two occupational areas within the low-complexity category: 18.1% for routine blue-collar jobs and 20.4% for routine clerical jobs. The average of 19.3% for low-complexity jobs is close to the Schmidt and Hunter (1983) mean figure of 20.0% for all the jobs they stud-

(*Text continues on page 36*)

Table 5
Incumbent Output Standard Deviations as Percentage of Mean Output (SD_p): New Studies (Except Sales)

Study	Occupation	N	Observed incumbent SD_p	Time period (weeks)	Reliability	Reliability corrected SD_p
		Low complexity				
Routine blue collar						
Vineberg & Taylor (1972)[a]	Armor crewman	374	18.3	JS	.78	16.2
U.S. Job Service (1966)[a]	Arc welder	49	18.1	JS	.78	16.0
Wyatt & Langdon (1932)[b]	Tile sizing & sorting	18	21.0	4	.83	19.1
Baumberger, Perry, & Martin (1921)[c]	Machine operator	76	14.7	0.5	.38	9.1
Hearnshaw (1937)[d]	Paper sorters	18	10.1	13	.74	8.7
Blum & Candee (1941)[b]	Package wrappers	27	26.5	4	.83	24.1
Blum & Candee (1941)[b]	Package packers	10	18.0	4	.83	16.4
No. studies and average		7	18.1			15.7
Routine clerical						
U.S. Job Service (1972)[a]	Proofreader	57	20.9	JS	.78	18.5
U.S. Job Service (1976)[a]	Grocery checker	92	21.9	JS	.78	19.3
U.S. Postal Service (1981)[e]	Mail carriers	374	23.2	JS	.94	22.5
U.S. Postal Service (1981)[f]	Mail handlers	373	26.7	JS	.72	22.7
Gael, Grant, & Ritchie (1975a)[a]	Telephone operator	1,091	20.0	JS	.78	17.7
Corts, Muldrow, & Outerbridge (1977)[e]	Customs inspector	188	17.6	JS	.80	15.7
Baumberger & Martin (1920)[e]	Telegraph operator	14	20.3	0.6	.42	13.2
Hay (1943)[d]	Machine bookkeepers	39	9.5	0.8	.78	8.4
Gaylord, Russell, Johnson, & Severin (1951)[b]	File clerks	61	18.2	26	.97	17.9
Maier & Verser (1982)[e]	Toll-ticket sorters	13	33.3	0.2	.20	14.9
Gael et al. (1975b)[a]	Clerks	402	22.4	JS	.78	19.8
No. studies and average		11	21.3			17.3
		Medium complexity				
Crafts						
Whipple, Baldin, Mager, & Vineberg (1969)						
Group 1[e]	Radar mechanics	107	47.8	JS	.71	40.3
Group 2[e]	Radar mechanics	51	23.8	JS	.71	20.1
Vineberg & Taylor (1972)[a]	Cook	364	22.3	JS	.78	19.7
Vineberg & Taylor (1972)[a]	Repairman	385	24.2	JS	.78	21.4
No. studies and average		4	29.5			25.4
Clerical with decision making						
U.S. Postal Service (1981)[e]	Mail distribution	417	40.0	JS	.96	39.2
Vineberg & Taylor (1972)[a]	Supply specialist	394	30.0	JS	.78	26.5
Trattner, Corts, van Rijn, & Outerbridge (1977)[e]	Claims authorizer	233	24.1	JS	.72	20.5
DeSimone, Alexander, & Cronshaw (1986)[b]	Claims evaluators	176	24.6	52	.99	24.5
Ledvinka, Simonet, Neiner, & Kruse (1983)[d]	Claims evaluators	15	30.9	4	.85	28.5
No. studies and average		5	29.9			27.8
		High complexity				
Professional judgment						
Campbell, Crooks, Mahoney, & Rock (1973)[e]	Cartographic technician	443	47.9	JS	.49	33.5
Altman & Weil, Inc. (1981)[b,g]	Attorneys (partners)	4,659	50.3	52	.99	50.0
Owens (1987)[b,g]	Physicians	7,567	52.8	52	.99	52.5
Theodore & Sutter (1967)[b,h]	Physicians	1,754	63.7	1	.55	47.2
American Dental Association (1969a)[b,g]	Dentists	4,023	51.4	52	.99	51.1
American Dental Association (1969b)[b,h]	Dentists	4,023	43.1	52	.99	42.9
No. studies and average		7	51.5			46.2

[a] R_{YY} was not given; the mean from Table 2 was used. [b] R_{YY} was not given; the mean from Table 1 for that time period was used. [c] R_{YY} was not given; the Spearman-Brown formula was used to compute partial week reliability from the mean reliability for 1 week (.55) in Table 1. [d] R_{YY} was given in the study for the correct time period. [e] Job sample R_{YY} was given in the study (see Table 2). [f] Average of the two job sample R_{YY}s was given in the study (see Table 2). [g] SD_p was computed from a national survey of yearly earnings. [h] SD_p was computed from a national survey of the number of patients seen and treated.

Table 6
Incumbent Output Standard Deviations as Percentages of Mean Output (SD_p): New Sales Studies

Study	Occupation	N	Observed incumbent SD_p	Time period (weeks)	Reliability	Reliability corrected SD_p
		Life insurance sales				
Wallace & Twichell (1953)[a]						
Group 1	Life insurance agent	140	53.1	100	.88	49.8
Group 2	Life insurance agent	112	64.4	100	.88	60.4
Brown (1981)[b]						
Company 1	Life insurance agent	3,590	136.0	52	.82	123.2
Company 2	Life insurance agent	768	116.1	52	.82	105.1
Company 3	Life insurance agent	949	105.1	52	.80	94.0
Company 4	Life insurance agent	1,606	132.1	52	.81	118.9
Company 5	Life insurance agent	752	96.8	52	.81	87.1
Company 6	Life insurance agent	893	115.1	52	.81	103.6
Company 7	Life insurance agent	606	122.2	52	.81	110.0
Company 8	Life insurance agent	793	130.1	52	.72	110.4
Company 9	Life insurance agent	771	140.0	52	.72	118.8
Company 10	Life insurance agent	658	129.3	52	.74	111.2
Company 11	Life insurance agent	561	114.0	52	.78	100.7
Company 12	Life insurance agent	406	130.7	52	.82	118.4
Bobko, Karren, & Parkington (1983)[c]	Insurance counselor	92	41.9	52	.80	37.5
No. studies and average		15	108.5			96.6
		Noninsurance sales				
Rush (1953)[b]	Office machine sales	100	48.2	41	.47	33.0
Bagozzi (1980)[d]	Industrial sales	122	29.0	52	.89	27.4
Burke & Frederick (1984)[d]	Sales manager	69	43.8	52	.89	41.3
Pearlman (1985)[d]	Sales account manager	42	80.9	52	.89	76.3
Greer & Cascio (1987)[d]	Soft drink route sales	62	35.3	52	.89	33.3
No. studies and average		5	47.4			42.3

[a] R_{YY}s were not given; the Spearman-Brown formula was used to compute R_{YY} for 100 weeks on the basis of the mean figure for life insurance agents for 52 weeks, from Table 3. [b] R_{YY}s were given in the study for the correct time period; subjects were agents in their first year on the job. [c] R_{YY} was not given; the estimate used was the mean figure for life insurance agents for 52 weeks, from Table 3. [d] R_{YY} was not given; the estimate used was the figure from Table 3 for "other sales" for 52 weeks.

Table 7
Average Incumbent and Applicant Output Standard Deviations as Percentages of Mean Output (SD_p) for Occupational Groups

Occupation	No. studies	Observed incumbent SD_p	Reliability corrected SD_p	Applicant SD_p
Low complexity				
Routine blue collar	22	18.4	14.6	18.1
Routine clerical	18	20.0	16.4	20.4
Average			15.5	19.3
Medium complexity				
Crafts	7	27.7	23.3	28.9
Decision-making clerical	5	29.9	27.8	34.5
Average			25.6	31.8
High complexity				
Professional judgment	7	51.53	46.2[a]	47.5[b]
Sales				
Life insurance	15	108.5	96.6	120.0
Noninsurance sales	7	46.4	38.5	47.7

[a] Figure reflects the average of the reliability-corrected figures from Table 5 for professional jobs. [b] SD_p figures for attorneys, physicians, and dentists were not corrected for range restriction because the samples used spanned the range of each profession and were close to representative of the potential applicant pool. SD_p figures for cartographic technicians were corrected for range restriction because the sample came from a single employer and therefore represented the incumbent group within an organization.

Table 8
Output Ratios Between Extremes of Applicant Groups: Top Versus Bottom 1% on Output and General Mental Ability

Category	SD_p	Output	Mental ability[a]
Job complexity			
Low	19.3	3.17	2.26
Medium	31.8	12.33	4.50
High	47.5	Not normal[b]	Not normal[b]
Sales			
Life insurance	120.0	Not normal[b]	Not normal[b]
Other sales	52.5	Not normal[b]	Not normal[b]

[a] Figures are based on estimated validity for general mental ability of .75; this is the average *true score* correlation between measures of general mental ability and performance on content-valid job-sample measures (from Hunter, 1986). Figures for *measures* of ability (as opposed to true scores) are given in the text. [b] Evidence presented in the text indicates that these distributions are not normal; thus, ratios are not computed.

ied. For the medium-complexity jobs, the two subdivisions are again fairly close: 28.9% for crafts and 34.5% for clerical jobs with decision-making components. The average value of 31.8% for medium-complexity jobs is considerably larger than the average for low-complexity jobs (19.3%), as expected. The average for high-complexity jobs (47.5%) is considerably higher than for medium-complexity jobs, again as expected. Mean applicant SD_p values for low-, medium-, and high-complexity jobs are approximately 19%, 32%, and 48%, respectively. Job complexity, measured in terms of the information-processing demands made by the job, is strongly related to the variability of output as a percentage of mean output. Furthermore, this relationship can be very large: The average coefficient of variation is almost 2.5 times larger for high-complexity jobs than for low-complexity jobs. This means that, other things being equal, selection utility per selectee (expressed as the percentage increase in output) will be nearly 2.5 times greater in high-complexity than in low-complexity jobs.

Life insurance sales jobs again yield extreme values. The mean applicant SD_p is 120.0%, by far the largest of all the means. Noninsurance sales show a mean value of 47.7% for applicants, which is similar to that for high-complexity jobs, even though these jobs are of only medium complexity. With respect to the impact of complexity on SD_p values for jobs in general, all sales jobs show "off-line" SD_p values, and the effect is most extreme for life insurance sales jobs. A possible explanation for this finding is discussed later.

Practical Implications of Variation in Output

The implications of the obtained SD_p values for personnel selection can be illustrated most straightforwardly by comparing extreme individuals. For a normal distribution, there is no upper or lower bound; more extreme values merely become less and less probable. However, essential ranges can be obtained by defining top and bottom categories, such as the top and bottom 1%. Table 8 was constructed assuming a normal distribution where possible (see later). Only applicant distributions were considered because these are the critical distributions for personnel selection. In low-complexity jobs, those in the top 1% on performance would average 2.67 standard deviations above the mean and would have a mean performance of 100 + 2.67(19.3) = 152, whereas those in the bottom 1% would have a mean of 100 − 2.67(19.3) = 48. The ratio of performance between the top and bottom 1% of workers would be 152/48 = 3.17. Thus, in low-complexity jobs, if people are hired randomly from the applicant pool, the top 1% of workers can be expected to produce 3 times as much as the bottom 1% of workers. If the output distribution is actually somewhat skewed rather than normal, this ratio would likely be larger.

Next, consider applicants for medium-complexity jobs, for which the output ratio is 31.8%. The mean output of those in the top 1% is 100 + 2.67(31.8) = 185% of average, whereas the mean for those in the bottom 1% is only 15% of average. Thus, extremely good workers can be expected to outperform extremely poor workers in medium-complexity jobs by a factor of 185/15 = 12.33. That is, in medium-complexity work, extremely good workers can be expected to outperform extremely poor workers by over 12 to 1. Again, if the output distribution is actually somewhat skewed rather than normal, this ratio would likely be even larger.

In high-complexity jobs, the estimate of the output standard deviation is 47.5%. If the distribution of performance were normal, the performance in the top 1% would be 100 + 2.67(47.5) = 227, whereas the performance in the bottom 1% would be 100 − 2.67(47.5) = −27. Whereas it is not unreasonable that top workers might be more than twice as productive as average workers, it is not likely that bottom workers would have negative output. It is more likely that the negative value arises from a nonnormal performance distribution in high-complexity jobs. If the distribution is positively skewed, then there may be no workers at 2 or more standard deviations below the mean. The data thus suggest that the low output point for high-complexity work is probably at or near zero, implying that in applicant populations, low performers cannot learn the job at all. For high-complexity work, the ratio of performance for high performers to low performers would be 227/0, which is meaningless (or "infinite" to the mathematician). An alternative approach to illustrating the extent of difference in high-complexity work is to compare high-performance workers with average workers. For typical positively skewed distributions, the top 1% averages even farther above the mean than 2.67 standard deviations. Thus, if we use the normal model used for lower complexity jobs, the ratio will be underestimated. This ratio is 227/100 for high-complexity jobs, 185/100 for medium-complexity jobs, and 152/100 for low-complexity jobs. Thus, in low-complexity jobs, the top 1% averages 52% more than the average employee. For medium-complexity jobs this figure is 85%, and for high-complexity jobs it is 127%.

Finally, consider sales jobs. As shown in Table 8, the mean applicant standard deviation for life insurance sales is 120.0%, over twice as large as the value for any other category examined here. For noninsurance sales, the mean SD_p is 47.7%, which is similar to the value for high-complexity jobs. Thus, both types of sales productivity are positively skewed. However, it may be that the high SD_p values for sales are not caused by levels of general mental ability that are insufficient for full mastery of the job. Instead, the skew may arise from a multiplicative effect of various traits and abilities on performance. A skewed distribution can be produced by multiplicative or combinatorial trait

requirements. For example, high sales performance may require high cognitive ability *and* a pleasant exterior personality *and* enough drive to close sales. If the distribution on each separate trait were normal, and if the traits multiply to predict sales, then the performance distribution would be highly skewed. For example, suppose that *high* and *low* on each trait is scored as 1 and 2, respectively. The product scores across the three traits would be 1, 2, 4, or 8, with probabilities 1/8, 3/8, 3/8, or 1/8, respectively. Mean performance would be 3.375 and the standard deviation would be 2.058. The lowest performance (level 1) is only 1.15 standard deviations below average, whereas highest performance (level 8) is 2.24 standard deviations above the mean. Thus, the distribution is quite skewed. In addition to sales jobs, a multiplicative process of this sort might also operate in high-complexity jobs, wherein output distributions also appear to be skewed.

Cognitive Ability Group Differences in Output

The data presented in this study show that individual differences in output are very large. It is clear that if people could be selected for jobs on the basis of a reliable measure of output, the differences in output between those selected and the average for the applicant pool would be very large. However, selection can rarely be based on output measures because output is not known prior to hiring. Instead, the employer must select on the basis of measures that have been shown to correlate with (and thus predict) future output. Because the validity of such measures is never perfect (i.e., validity is always less than $r_{xy} = 1.00$), the differences in output between the selected group and the average for the applicant pool will not be as large as for selection on a measure of output itself. However, these differences can still be substantial. One measure that has been shown to be correlated with performance on virtually all jobs is general mental ability (Hunter, 1980, 1986; Hunter & Hunter, 1984). This section examines output differences between the highest and lowest applicant groups on general mental ability. Actual calculations are carried out only for medium-complexity, white-collar work, although similar results apply to other job classifications as well. The cumulative research literature is used to compute the percentage difference in work output between extreme cognitive ability groups. Cumulative research has shown that performance is linearly related to general cognitive ability (Schmidt et al., 1979). Thus, differences in output between cognitive ability groups can be computed by using linear regression.

Because general cognitive ability is measured on interval rather than ratio scales, the applicant mean and standard deviation are arbitrary; we use a mean of 100 and a standard deviation of 15. If ratio scale measurement of performance is scaled so that mean applicant output is 100, the output standard deviation is then the applicant output standard deviation ratio (SD_p), which varies from one complexity level to another. For medium-complexity, white-collar work, the empirical value in Table 7 is 34.5%. The ability output correlation is the average validity of cognitive ability for predicting work sample performance attenuated to the reliability of whatever test is used. If a test of perfect reliability were used, the applicant correlation would be .75 (Hunter, 1986; see also Hunter, 1983). We use this figure initially here because our concern at this point is with actual ability rather than imperfect *measures* of ability.

The slope and intercept of the regression line are given by

$$b = \text{slope} = r_{XY}S_Y/S_X = .75(34.5)/15 = 1.725$$

$$a = \text{intercept} = m_Y - bm_X = 100 - 1.725(100) = -72.5.$$

The negative intercept is not meaningful because the cognitive test is scored to have a mean of 100 and a standard deviation of 15. Thus, 99.9% of applicants have scores higher than 55. The regression equation is $Y = 1.725X - 72.5$.

The bottom 1% of the ability distribution has an average score of 60; the top 1% averages 140 on the test. Based on our regression equation, the mean output values for these groups are 31.0 and 169.0. Thus, the ratio of output for the extreme ability groups is 169/31 = 5.45. That is, the highest cognitive ability applicants would outperform the lowest cognitive ability applicants by over 5 to 1, a very large difference in performance. However, this difference assumes a perfectly reliable measure of cognitive ability. Ordinarily, the reliability of the measure of cognitive ability would be approximately .80. At this level of reliability, the output ratio of the top 1% of scores to the bottom 1% would be 3.46 to 1. This ratio is lower, but it is clear that it is still substantial and that there is considerable room for improvement in performance due to good personnel selection if the organization can be selective in its hiring.

Finally, all the ratios of highest to lowest output employees presented in this section as examples would be larger if more extreme groups (e.g., top and bottom 0.5%) were used. Conversely, all ratios would have been somewhat smaller if less extreme groups had been used (e.g., top and bottom 5% or 10%). We used the top and bottom 1% because we judged that these groups came closest to most people's conception of extreme groups. This article presents the information needed by the reader to compute output ratios for any definition of extreme output groups. For example, a reader might be interested in the ratio of output of the top 50% to the bottom 50%. By using the data in Table 7, one can compute this ratio for either applicants or incumbents and for any of the job categories presented. Space considerations prevent a more detailed presentation of such ratios in this article.

Implications for Theory Development

The findings of this research contribute to the foundation for a theory of work performance that has job complexity as one of its central constructs. For example, other factors (such as validity and selection ratio) being equal, the findings of this research indicate that the percentage increases in output produced by improved selection are about 2.5 times greater in high-complexity jobs than in low-complexity jobs (47.5/19.3 = 2.46).

The explanation for the strong relation between complexity of information-processing requirements and SD_p is an interesting question but one that cannot be definitively answered at present. The obvious hypothesis is that increasing complexity *causes* increasing values of SD_p. In our judgment, this hypothesis is far more plausible than its opposite, which is the hypothesis that SD_p levels cause complexity. However, the hypothesis that information-processing complexity, as measured in this study, causes SD_p does not appear to fit the data for sales jobs (see preceding discussion) in the same manner as it fits the data from other jobs. Other constructs may be required for sales jobs.

In addition, it is logically possible that other variables associated with complexity (such as increased employee discretion or autonomy) may have a causal impact on SD_p. Other key constructs in a theory of job performance would be general mental ability, job experience, job knowledge (Hunter, 1980, 1986; Schmidt, Hunter, Outerbridge, & Goff, 1988), and psychomotor ability (Hunter, 1980). We are currently working to develop such a theory. One question that must be addressed in the development of this theory is the relative construct validity of job sample measures and supervisory ratings as measures of job performance (see Appendix B).

Implications for Labor Economics

In addition to the above implications, it now appears that research findings on SD_p may have implications in the related field of labor economics. The theory of efficient labor markets (ELM), central to labor economics, postulates that to the extent that valid information on individual output is freely available, employers will compensate individual employees in proportion to their performance or output (Frank, 1984). Recent research in industrial/organizational psychology in connection with selection utility has resulted in a collection of evidence indicating that there are wide variations in output among employees who are paid identical or very similar wages (e.g., Schmidt & Hunter, 1983). Although other interpretations are possible (Frank, 1984) within the context of ELM theory, this finding suggests that employer information on the output of individual employees may be limited in accuracy (Bishop, 1987a). The theory of ELM also predicts that employers making hiring decisions will use all available valid information to predict later performance on the job. Yet, many employers fail to use valid performance predictors, such as measures of general mental ability. According to ELM theory, this fact would suggest either that employers are unaware of such valid performance predictors or that the perceived costs (e.g., in terms of potential litigation) of using such information is greater than the perceived benefits in increased output. (Of course, these perceptions may be erroneous.) The labor economist who seems to have taken the lead to date in exploring these implications is Bishop (1987a, 1987b), but others have also contributed (e.g., Mueser & Mahoney, 1987). Although this work is very recent and is still controversial in labor economics, in our judgment it is a positive development that research findings in our field have potential implications for theoretical and empirical work in another social science. By providing more accurate and differentiated values for SD_p, our study may contribute further to this development.

Other Implications

In addition to the finding that SD_p increases with increases in job complexity, this study makes several other contributions. First, it summarizes the information in the literature on the reliability of output measures, job sample measures, and sales measures; to our knowledge, this information is not available elsewhere. A noteworthy finding is that the reliability of counts of actual output appears to be lower than may generally be believed. Second, it explains and illustrates the effect of unreliability of (ratio scale) criterion measures on SD_p estimates and demonstrates the appropriate corrections. Third, it provides evidence that SD_p values for sales jobs are quite different from those for nonsales jobs at the same levels of complexity; such evidence is not systematically presented elsewhere in the literature. Fourth, it presents evidence for the economic importance of individual differences in job performance without using estimates of SD_y; estimates of SD_y (the dollar value of the standard deviation of job performance) must, unlike SD_p, be estimated judgmentally. Finally, the findings of this study help to lay the foundation for a theory of work performance, as noted earlier.

Summary and Conclusions

Good personnel selection can produce increased productivity (output) only to the extent that there are large individual differences in performance. This study analyzed data on the extent of individual differences in productivity (output), based on 68 studies measuring work output on the job and 17 work sample studies with ratio scale measurement. In each study, the ratio of the standard deviation in output to mean output was calculated. This ratio times 100 is the standard deviation of output measured as a percentage of average output (SD_p).

Schmidt and Hunter (1983) concluded that SD_p is at least 20% of mean output for incumbents. This study refines and elaborates on that estimate. For incumbents in routine clerical or blue-collar work, the output standard deviation ratio was found to be closer to 15% than 20%. However, in higher complexity jobs, the output standard deviation ratios are larger. For incumbents in medium-complexity jobs, the standard deviation ratio was found to be about 25% rather than 20%. For incumbent workers in high-complexity jobs, the output standard was 46% rather than 20%. For sales, the incumbent output standard deviation ratio is very large (97%) for life insurance sales, and about 39% for other sales jobs.

But job incumbents are not representative of applicants. Analysis of the data from 515 U.S. Employment Service studies showed that on a measure of general cognitive ability, the observed score standard deviation of incumbents averages only 71% of the standard deviation of applicants (i.e., $u = .71$). Appendix B demonstrates that the u value for ability true scores (actual ability) is even smaller (.61) and that performance standard deviations are attenuated on average by at least 20%. Thus, the applicant performance standard deviations needed for personnel selection utility equations are at least 24% higher on average than the performance standard deviations for incumbents. Moving from routine- to medium-complexity to professional work, the output mean standard deviation for applicants was found to vary from 19.3% to 31.8% to 47.5%, respectively. For life insurance sales applicants, the average standard deviation was 120.0%, and for other sales the average was 47.7%.

Individual differences in work output are very large. For medium-complexity work, for example, extreme 1% performance groups differ by a factor of 12 to 1, and extreme 1% cognitive ability groups differ by a factor of over 5 to 1. Using cognitive ability measures with a reliability of .80, this ratio is still approximately 3.5 to 1. Thus, there are large gains in productivity to be made by selecting better workers. In addition to these important practical implications, the finding that output variability relative to mean output increases with job complexity may prove to have broader and more theoretical implications in industrial/organizational psychology, because individual differ-

ences in job performance are the key dependent variable in this area of applied differential psychology. Finally, it now appears that research findings on individual differences in work output may have implications for theoretical and empirical work in the sister social science of labor economics.

There is a need for more research on the variability of employee output by complexity level of job. The need for additional data is greatest for high-complexity jobs, but additional data are also needed for medium-complexity jobs. As more data become available, the estimates of SD_p can be made more precise. In addition, researchers should give more attention to the question of the reliability of output and sales data. As illustrated in this study, many researchers do not report the reliability of output data, necessitating extrapolation of reliabilities from other studies. Finally, the data reported in this study indicate that the reliability of counts of actual employee output is lower than might generally have been believed. This finding points to a need for more research on the conditions that might affect the reliability and stability of employee output over time.

References

Alexander, R. A., Carson, K. P., Alliger, G. M., & Cronshaw, S. F. (1989). Empirical distributions of range restricted SD_x in validity studies. *Journal of Applied Psychology, 74,* 253–258.

Altman & Weil, Inc. (1981). *The 1981 survey of law firm economics.* Ardmore, PA: Author.

American Dental Association: Bureau of Economic Research and Statistics (1969a). 1968 survey of dental practice II: Income of dentists by location, age, and other factors. *Journal of the American Dental Association, 78,* 342–346.

American Dental Association: Bureau of Economic Research and Statistics (1969b). 1968 survey of dental practice VII: Number of patients and patient visits. *Journal of the American Dental Association, 79,* 378–380.

Bagozzi, R. P. (1980). Performance and satisfaction in an industrial sales force: An examination of their antecedents and simultaneity. *Journal of Marketing, 44,* 65–77.

Barnes, R. M. (1958). *Time and motion study* (4th ed.). New York: Wiley.

Baumberger, J. P., & Martin, E. G. (1920). Fatigue and efficiency of smokers in a strenuous mental occupation. *Journal of Industrial Hygiene, 2,* 207–214.

Baumberger, J. P., Perry, E. E., & Martin, E. G. (1921). An output study of users and non-users of tobacco in a strenuous physical occupation. *Journal of Industrial Hygiene, 3,* 1–10.

Bishop, J. (1987a). The recognition and reward of employee performance. *Journal of Labor Economics, 5*(No. 4, Pt. 2), S36–S56.

Bishop, J. (1987b). *Information externalities and the social payoff to academic achievement* (Paper No. 87-06). Center for Advanced Human Resources Studies, Cornell University, Ithaca, NY.

Blum, M., & Candee, B. (1941). The selection of department store packers and wrappers with the aid of certain psychological tests. *Journal of Applied Psychology, 25,* 76–85.

Bobko, P., Karren, R., & Parkington, J. J. (1983). Estimation of standard deviations in utility analysis: An empirical test. *Journal of Applied Psychology, 68,* 170–176.

Brown, S. H. (1981). Validity generalization and situational moderation in the life insurance industry. *Journal of Applied Psychology, 66,* 664–670.

Bureau of National Affairs, Inc. (1983, May 5). *BNA Bulletin to Management* (No. 1727). Washington, DC: Author.

Burke, M. J., & Frederick, J. T. (1984). Two modified procedures for estimating standard deviations in utility analysis. *Journal of Applied Psychology, 69,* 482–489.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in the prediction of job performance: a six-year study.* Princeton, NJ: Educational Testing Service.

Ciocco, A., & Altman, I. (1943). The patient loads of physicians in private practice. *Public Health Reports, 58,* 1329–1351.

Corts, D. B., Muldrow, T. W., & Outerbridge, A. M. (1977). *Research base for Examination (PACE): Prediction of job performance for customs inspectors.* Washington, DC: U.S. Office of Personnel Management.

DeSimone, R. L., Alexander, R. A., & Cronshaw, S. F. (1986). Accuracy and reliability of SD_y estimates in utility analysis. *Journal of Occupational Psychology, 59,* 93–102.

Evans, D. W. (1940). Individual productivity differences. *Monthly Labor Review, 50,* 338–341.

Frank, R. H. (1984). Are workers paid their marginal products? *American Economic Review, 74,* 549–571.

Gael, S., Grant, D. L., & Ritchie, R. J. (1975a). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology, 60,* 411–419.

Gael, S., Grant, D. L., & Ritchie, R. J. (1975b). Employment test validation for minority and nonminority clerks with work sample criteria. *Journal of Applied Psychology, 60,* 420–426.

Gaylord, R. H., Russell, E., Johnson, C., & Severin, D. (1951). The relation of ratings to production records: An empirical study. *Personnel Psychology, 4,* 363–371.

Greer, O. L., & Cascio, W. F. (1987). Is cost accounting the answer? Comparison of two behaviorally based methods for estimating the standard deviation of job performance in dollars with a cost-accounting approach. *Journal of Applied Psychology, 72,* 588–595.

Hay, E. N. (1943). Predicting success in machine bookkeeping. *Journal of Applied Psychology, 27,* 483–493.

Hearnshaw, L. S. (1937). Selection tests for paper sorters. *Journal of Occupational Psychology, 11,* 145–153.

Hunter, J. E. (1980). *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, DC: U.S. Employment Service.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement theory* (pp. 257–266). Hillsdale, NJ: Erlbaum.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340–362.

Hunter, J. E., & Hirsh, H. R. (1987). Applications of meta-analysis. In C. L. Cooper & I. T. Robertson (Eds.), *Review of industrial psychology* (Vol. 2, pp. 321–357). New York: Wiley.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Hunter, J. E., & Schmidt, F. L. (1983). Quantifying the effects of psychological interventions on employee job performance and work force productivity. *American Psychologist, 38,* 473–478.

Kahn, D. F., & Hadley, J. M. (1949). Factors related to life insurance selling. *Journal of Applied Psychology, 33,* 132–140.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 65,* 507–516.

Klemmer, E. T., & Lockhead, G. R. (1962). Productivity and errors in two keying tasks: A field study. *Journal of Applied Psychology, 46,* 401–408.

Lawshe, C. H. (1948). *Principles of personnel tests.* New York: McGraw-Hill.

Ledvinka, J., Simonet, J. K., Neiner, A. G., & Kruse, B. (1983). *The dollar value of JEPS at Life of Georgia.* Unpublished technical report.

Maier, N. R. F., & Verser, G. C. (1982). *Psychology in industrial organizations* (5th ed.). Boston: Houghton Mifflin.

Manson, G. E. (1925). What can the application blank tell? Evaluation of items in personal history records of four thousand life insurance salesmen. *Journal of Personnel Research, 4,* 73–99.

McCormick, E. J., & Tiffin, J. (1974). *Industrial psychology.* Englewood Cliffs, NJ: Prentice-Hall.

McEvoy, G. M., & Cascio, W. F. (1986). *A meta-analysis of the relationship between employee performance and turnover.* Unpublished manuscript, Utah State University.

Mueser, P., & Mahoney, T. (1987). *Cognitive ability, human capital and employer screening: Reconciling labor market behavior with studies of employee productivity.* Unpublished manuscript, University of Missouri—Columbia.

Owens, A. (1987). Doctors' earnings: On the rise again. *Medical Economics, 64,* 212–237.

Pearlman, K. (1985, August). *Development of a dollar criterion for high-level sales jobs.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles, California.

Rimland, B., & Larson, G. E. (1986). Individual differences: An underdeveloped opportunity for military psychology. *Journal of Applied Social Psychology, 16,* 565–575.

Rothe, H. F. (1946). Output rates among butter wrappers: II. Frequency distributions and an hypothesis regarding the "restriction of output." *Journal of Applied Psychology, 30,* 320–327.

Rothe, H. F. (1947). Output rates among machine operators: I. Distributions and their reliability. *Journal of Applied Psychology, 31,* 484–489.

Rothe, H. F. (1970). Output rates among welders: Productivity and consistency following removal of a financial incentive system. *Journal of Applied Psychology, 54,* 549–551.

Rothe, H. F. (1978). Output rates among industrial employees. *Journal of Applied Psychology, 63,* 40–46.

Rothe, H. F., & Nye, C. T. (1958). Output rates among coil winders. *Journal of Applied Psychology, 42,* 182–186.

Rothe, H. F., & Nye, C. T. (1961). Output rates among machine operators: III. A nonincentive situation in two levels of business activity. *Journal of Applied Psychology, 45,* 50–54.

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

Schmidt, F. L., & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology, 68,* 407–414.

Schmidt, F. L., Hunter, J. E., McKenzie, R., & Muldrow, T. (1979). The impact of valid selection procedures on workforce productivity. *Journal of Applied Psychology, 64,* 609–626.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Goff, S. (1988). Joint relation of experience and ability with job performance: Test of three hypotheses. *Journal of Applied Psychology, 73,* 46–57.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Trattner, M. H. (1986). The impact of job selection methods on size, productivity and payroll cost of the federal workforce: An empirically based demonstration. *Personnel Psychology, 39,* 1–29.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on workforce productivity. *Personnel Psychology, 35,* 333–347.

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. park ranger for three modes of test use. *Journal of Applied Psychology, 69,* 490–497.

Stead, W. H., & Shartle, C. L. (1940). *Occupational counseling techniques.* New York: American Book.

Strong, E. K. (1935). Interests and sales ability. *Personnel Journal, 13,* 204–216.

Theodore, C. N., & Sutter, G. E. (1967). A report on the first periodic survey of physicians. *Journal of the American Medical Association, 202,* 180–189.

Tiffin, J. (1947). *Industrial psychology* (2nd ed.). New York: Prentice-Hall.

Tiffin, J., & McCormick, E. J. (1965). *Industrial psychology* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Trattner, M. H., Corts, D. B., van Rijn, P. P., & Outerbridge, A. M. (1977). *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for claims authorizers in the social insurance claims examining occupation.* Washington, DC: U.S. Office of Personnel Management.

U.S. Department of Labor, Employment and Training Administration (1977). *Dictionary of occupational titles* (4th ed.). Washington, DC: U.S. Government Printing Office.

U.S. Job Service (1966). *Development of USES aptitude test battery for arc welder 810.884.* Washington, DC: U.S. Employment Service.

U.S. Job Service (1972). *Development of USES aptitude test battery for copy holder 209.588 and proofreader 209.688.* Washington, DC: U.S. Employment Service.

U.S. Job Service (1976). *Development of USES aptitude test battery for grocery checker 229.468.* Washington, DC: U.S. Employment Service.

U.S. Postal Service (1981). *Validation report for positions of city carrier, mail handler, distribution clerk, and distribution clerk machine.* Washington, DC: United States Postal Service, Employment and Placement Division.

Validity Information Exchange No. 11-27 (1958). *Personnel Psychology, 11,* 583.

Vineberg, R., & Taylor, E. N. (1972). *Performance in four army jobs by men at different aptitude (AFQT) levels.* Alexandria, VA: Human Resources Research Organization.

Wallace, S. R., & Twichell, C. (1953). An evaluation of a training course for life insurance agents. *Personnel Psychology, 6,* 25–43.

Wechsler, D. (1952). *Range of human capacities* (2nd ed.). Baltimore, MD: Williams and Wilkins.

Weekley, J. A., & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology, 72,* 484–487.

Whipple, J. W., Baldin, R. D., Mager, R. F., & Vineberg, R. (1969). A three-hour performance test to evaluate job effectiveness of Army radar mechanics. Alexandria, VA: Human Resources Research Organization.

Wyatt, S., & Langdon, J. N. (1932). Inspection processes in industry. In *Industrial Health Research Board Report* (No. 63). London, Great Britain: His Majesty's Stationery Office.

Appendix A

Effect of Location on SD_p

One reviewer was concerned that there may be differences in mean earnings between areas within regions and that these differences could inflate SD_p values for dentists and medical doctors. There are two considerations that suggest that this could not occur. First, this argument would not appear to apply to SD_p values based on number of patients treated. Consider the following figures from Table 5:

Physicians

SD_p based on earnings = 52.8

SD_p based on patients treated = 63.7

SD_p based on patients treated is 21% larger.

Dentists

SD_p based on earnings = 51.4

SD_p based on patients treated = 43.1

SD_p based on patients treated is 16% smaller.

Combined

SD_p based on earnings = 52.1

SD_p based on patients treated = 53.4

These figures indicate that there is no systematic tendency for SD_p values based on earnings to be larger than SD_p values based on number of patients treated. On the average these two methods yield virtually identical figures: 52.1 versus 53.4. Thus, it does not appear that SD_p values based on earnings are inflated. Second, we located additional data that indicate that SD_p is not smaller in more circumscribed locations (Ciocco & Altman, 1943). For example, for physicians in the prime earning years of 45–64, SD_p by location was

District of Columbia	.76
Baltimore	.80
Maryland (excluding Baltimore)	.61
Georgia (urban)	.63
Georgia (rural)	.62

Even in rural Georgia, the figure is larger than the average in our Table 5 (52.8 and 63.7; average = 58.25). Thus, controlling for location does not appear to lead to smaller SD_p values.

Appendix B

Applicant Standard Deviations

This appendix derives an estimate of the ratio of performance standard deviations for applicants and incumbents. This derivation indicates that applicant output standard deviations are at least 24% larger than incumbent output standard deviations. However, this value represents an underestimate of the extent of restriction in performance, because it considers only the effect of restriction on performance caused by restriction on general cognitive ability due to hiring practices. It ignores direct restriction on the performance distribution caused by differential attrition. That is, this calculation ignores the facts that low-performing workers (a) are more likely to be fired and (b) quit earlier on their own (Hunter & Hirsh, 1987, reanalyzing the data from McEvoy & Cascio, 1986). It also ignores the fact that high-performing workers are often promoted out of the job in question into other, higher level jobs, such as supervision, further increasing range restriction.

Restriction in Range on Cognitive Ability

In the process of validating the General Aptitude Test Battery (GATB), the U.S. Employment Service conducted 415 validation studies on a sample of jobs spanning the job spectrum. Hunter (1980) calculated the extent of restriction in range for the GATB measure of general cognitive ability (u) and found it to be about .67 in all job families. Later, when more information on range restriction in this data set became available, Alexander, Carson, Alliger, and Cronshaw (1989) recalculated this average as .71. This latter figure is the one used here. However, this figure expresses the extent of restriction on test scores; the degree of restriction on actual ability is higher. The first step in this Appendix calculates the range restriction ratio for general cognitive ability itself.

Applicant population. Hunter (1980) performed an analysis of reliability (generalizability) on GATB aptitude and ability measures for an applicant population and found the coefficient of generalizability for the general cognitive ability measure to be .80. Let us denote the variance of actual ability, test score (the *measure* of ability), and error by VA, VX, and VE, respectively. From the theory of reliability, we know that

$$\text{Reliability} = .80 = \text{VA}/\text{VX}.$$

From the fact that

$$\text{VX} = \text{VA} + \text{VE},$$

we have

$$\text{VE} = \text{VX} - \text{VA}.$$

If we scale the test so that the applicant standard deviation is 1.00, then VX = 1.00 and hence

$$\text{VA} = \text{VA}/\text{VX} = .80$$

$$\text{VE} = \text{VX} - \text{VA} = 1.00 - .80 = .20.$$

Incumbent population. Let us add a prime to denote the variances in the incumbent population: VX′, VA′, and VE′. Because the process of measurement is the same in both populations, the error variance will not change. Thus, we have

$$\text{VE}' = \text{VE} = .20.$$

By definition, the range restriction coefficient is the ratio of test variances. Thus,

$$\text{VX}'/\text{VX} = u^2 = (.71)^2 = .50.$$

Because the test is scaled so that VX = 1.00, we have

$$\text{VX}' = .50$$

$$VA' = VX' - VE' = .50 - .20 = .30.$$

Thus, the square of the range restriction ratio for ability is

$$u^2 = VA'/VA = .30/.80 = .375.$$

That is, the extent of range restriction on general cognitive ability is

$$u = (.375)^{1/2} = .61.$$

Implied Output Range Restriction

Hunter (1986; see also Hunter, 1983), using meta-analysis, found the average applicant pool, true-score correlation between general cognitive ability and job performance measured by content-valid, work-sample performance measures in civilian work to be .75. Because we are primarily interested in the civilian economy, we use this figure. Approximately 97% of the U.S. work force is in the civilian economy. The corresponding true-score correlation for military jobs is considerably smaller (.53) and would yield somewhat smaller estimates of applicant pool SD_p values. The appropriately weighted average of the two correlations is .7434, essentially identical to the .75 value for civilian jobs. Schmidt et al. (1979) reviewed research showing the relationship between ability and performance to be linear. From these facts, we can derive the extent of performance restriction implied by restriction in range on ability. This analysis assumes that restriction on performance is solely indirect, that is, is due entirely to restriction on ability. Direct restriction on job performance (e.g., through promotion or termination) is ignored. If the only output restriction is that due to ability restriction, then the ratio of the applicant standard deviation to the incumbent standard deviation will be shown to be 1.24.

Applicant population. If ability true scores in the applicant population are in standard score form, then the regression of performance onto ability is given by

$$P = .75A + e,$$

where P is performance, A is general cognitive ability, and e is error of prediction. We have

$$VP = (.75)^2VA + Ve$$

$$1.00 = .5625(1.00) + Ve$$

$$Ve = 1.00 - .5625 = .4375.$$

Incumbent population. In this section, applicant population ability is expressed in standard score form. Thus, incumbent ability variance is the range restriction ratio .30/.80 = .375. For the incumbent population, we have

$$VP' = (.75)^2VA' + Ve$$

$$= .5625(.30/.80) + .4375 = .6484$$

The ratio of the standard deviations is $(.6484/1.00)^{1/2} = .805$. That is, the incumbent standard deviation is only 80.5% as large as the applicant standard deviation. That is, if applicants were hired at random, the performance standard deviation would be 1.00/.805 = 1.242, or 24% larger than the incumbent standard deviation observed in studies.

This figure is based on the average true-score correlation of .75 between general mental ability and content-valid job sample measures (Hunter, 1983, 1986). In our judgment, content-valid job-sample measures have greater construct validity as measures of job performance than supervisory ratings (Hunter, 1983), particularly when the job performance construct of interest is production or output, as is the case here. For this reason we did not use validities based on supervisory ratings. On the basis of Hunter's (1980) findings, the true-score correlations for general mental ability and supervisory ratings are .65 for high-complexity jobs, .57 for medium-complexity jobs, and .44 for low-complexity jobs (when complexity is as defined in the text and in Footnote 1). These values would lead to somewhat smaller corrections of incumbent SD_p values for range restriction. For example, the correlation of .57 for medium-complexity jobs yields a range correction factor of 1.12 (vs. our value of 1.24). For medium-complexity, white-collar jobs, estimated applicant pool SD_p would be 31.1 (vs. our value of 34.5). However, we believe these estimates of applicant SD_p values would be less accurate than those presented in Table 7 in the text.

Finally, we address a concern raised by a reviewer. The calculations in this appendix are based on the assumption that the average level of range restriction on general mental ability in the studies in Tables 4, 5, and 6 is approximately equal to the average level of range restriction found by Alexander et al. (1989) for the 415 widely varying GATB validity studies. In light of the fact that the number of studies is fairly large in both cases, and the fact that the studies were unselected with respect to range restriction, this assumption appears to be plausible. This assumption allows estimation of applicant pool SD_p values. In the case of any organization using the applicant pool SD_p estimates from this study, there is no necessity to assume that range restriction values in that organization are the same as the u = .71 GATB value (or are the same as those in studies in Tables 4, 5, and 6). It is not psychometrically relevant what the level of range restriction is in any organization that might use our applicant pool SD_p values (e.g., in a selection utility study). This is because the *incumbent* SD_p values in such an organization (which will be affected by level of range restriction) have no necessary relation to the *applicant pool* SD_p values for that organization (which are *not* affected by the level of range restriction *within* the organization). However, our estimates of applicant pool SD_p values are estimates of *averages*. Therefore, for reasons unrelated to range restriction within the organization, applicant pool SD_p values might be somewhat larger or smaller for particular organizations. However, short of estimating the incumbent SD_p value for that organization and then correcting this value based on the degree of range restriction in that organization, there would be no way to determine this. Such estimation will rarely be possible and therefore the estimates of applicant pool SD_p presented in Table 7 will usually be the most accurate available for any organization.

Received July 21, 1988
Revision received April 20, 1989
Accepted July 19, 1989 ■