The Correlation between Long-Term Productivity and Short-Term Performance Ratings of Harvester Operators

Thomas Purfürst, Ola Lindroos

Abstract – Nacrtak

Human operators are key determinants of the performance of most production systems, so individual performance is of intrinsic interest when evaluating current and proposed systems for forest operations. Such evaluations can be useful for diverse purposes, for instance, planning, incentive-setting, control and costing. Hence, various evaluation methods have been developed, all with pros and cons. Here, we compare subjective, short-term ratings of the work-related behavior of 12 harvester operators and their long-term output (harvested volume per unit time), based on observation periods of a few hours and data gathered over two months, respectively. It was found that competent raters can filter the many, interacting behavioral components and translate short-term observations into grades that reflect the operator's long-term output well (Spearman's $r_s > 0.9$). Moreover, substantial variations in performance values obtained by both methods were found, probably at least partly attributable to variations in individual performance of both the operators and the raters. We argue that both of the studied methods could be used to adjust population norms (e.g. productivity functions) to the individual's performance, with sufficient accuracy for normal production purposes (e.g. planning). However, in a scientific context it could be questioned whether the expected uncontrolled variation in operators' performance is most efficiently minimized by the introduction of uncontrolled variation in rater's behavior and/or historical data, or if other precautions could be taken to improve the reliability of the data.

Keywords: performance measurement, operator rating, CTL harvester thinning, StanForD, harvester operator, operator influence, human factors

1. Introduction – *Uvod*

The scientific discipline of work science systematically evaluates current and proposed human-machine-environment (HME) systems (Björheden 1991; Wilson 1998), with an intrinsic interest in the performance of systems. However, it is important to realize the ambiguity of the term performance, because it has different meanings depending on context. Interested readers are advised to compare, for instance, organizational-oriented literature (e.g. Tangen 2003; Tangen 2005) and literature oriented towards the focus of this paper; the performance of individuals. Since human operators are key determinants of the performance of most production systems, individual performance has been intensively investigated by, *inter alia*, work and organizational psychologists (Sonnestag and Frese 2002). Progress in recent decades has clarified and extended the concept of individual performance, which is now generally considered to be multi-dimensional and dynamic. In fact, methodologies to manage the extreme variations in individual performance that can occur have been proposed (Beheshti and Lollar 2008). Extensive models have also been proposed for exploring the effects of diverse aspects of individual performance. A thorough review of such models is beyond the scope of this paper, but numerous other authors have reviewed them (e.g. Arvey and Murphy 1998; Sonnestag and Frese 2002; Newman et al. 2004; Tubré et al. 2006).

However, to set the context for the study presented here the conceptual framework is roughly outlined below.

Two aspects of individual performance are generally differentiated: action and outcome, i.e. the behavior and results of individuals' actions, respectively (e.g. Arvey and Murphy 1998; Sonnestag and Frese 2002; Newman et al. 2004). Behavior, here, refers to what an individual does in the work situation (e.g. operating a harvester), but not all behavior performed during work is considered within the concept. As with the concept of forest work time differentiation (e.g. Björheden 1991), only behavior that is relevant to the operational goals is considered. Hence, the behavioral aspect of performance is not defined by the action itself, but by judgmental and evaluative processes (Sonnestag and Frese 2002). The outcome aspect refers to work-related results of the individual's behavior (e.g. harvested volume per unit time). The behavioral and outcome aspects are often related, but outcome depends on external factors in addition to the individual's behavior, and interactions between behavior and external factors may either enhance or adversely affect individuals' performance output. The influence of external factors can be readily understood in forestry, because of the strong influences of environmental factors (e.g. tree size and forwarding distance). Moreover, output is known to be highly dynamic (Vöry 1954; Steinlin 1955; Appelroth 1980) and there are substantial differences in output performance between operators; the most productive operators have been found to be at least 114%, 300% and 80% more productive than the least productive operators in manual (Harstela 1975), motor-manual (Reichel 1999) and mechanized work (Purfürst 2009), respectively. However, despite these well-known variations, individual performance is seldom considered in scientific studies. For example, in a review of 53 productivity models for harvester work, Purfürst (2009, p 24) found that 15 only provided information on the experience of the studied operators, three provided some additional information about the operators, and the remaining 35 models did not recognize the operator as an influential factor at all.

The motives for analyzing individual performance depend on whether the objectives are practical or scientific. For a practitioner, individual performance is something that should ideally be enhanced and optimized (Sonnestag and Frese 2002). In contrast, for a forest work scientist, the objective has usually been to normalize the operator influence in order to make valid generalizations for a population of operators. Hence, operator influence has mainly attracted interest as a noise factor. Two approaches for handling inevitable operator effects have been adopted in forest work science (Lindroos 2010). In one, output data are corrected to account for between-operator variations, either objectively or subjectively. Typically, if this is done subjectively, operator behavior (e.g. speed of movements) during the observed task is rated in relation to some kind of norm. Hence, the subjective method is often called performance rating, although it is a very specific application of performance rating, as discussed below. In the other approach, in which output data are not corrected to account for individual behavior, operator blocking is generally applied, i.e. during the observations all operators work with all tested methods/machines (Lindroos 2010).

Subjective correction of output data has a tradition in forestry in continental Europe and Great Britain, but it has been criticized for its subjectivity, i.e. it is biased by the expert's interpretations even when collecting the data (Mattson Mårn 1953; Steinlin 1955; Kärkkäinen 1975; Samset 1990; Samset 1992), whereas operator-blocking has been criticized for neglecting the well-known variability in human physiological and psychological characteristics (Appelroth 1989; Thompson 1992). A recent contribution that explored the strengths and limitations of operator-blocking (Lindroos 2010) and this work on the uses of performance rating in an updated framework, may (to some extent at least) mediate in the old, and sometimes heated, argument between advocates of the two approaches (see e.g. Sundberg 1988).

In disciplines concerned with the management of human resources, contemporary use of performance assessments is not mainly for minimizing operator influence, but for addressing the underlying factors that contribute to differences in individual performance. Hence, the interest lies in detecting, quantifying and analyzing both behavioral and output aspects of individual performance. In this context, subjective performance ratings by observers play a key role. The strengths and weaknesses of the many set-ups that can be used in performance ratings for these purposes have been scrutinized and discussed at length by work psychologists (e.g. Arvey and Murphy 1998; Sonnestag and Frese 2002; Newman et al. 2004), but surprisingly little attention has been paid to them in forestry work analyses. Some efforts have been made to evaluate and develop methodology for studies concerned with enhancing individual performance (Gellerstedt 2002; Ovaskainen et al. 2004; Ovaskainen 2005; Ovaskainen and Heikkilä 2007), but there has been little corresponding focus on methods for assessing individual performance.

Commonly used methods that enable assessment of (individual) performance in forestry are time studies, performance rating and follow-up studies (analyses of historical output records) (Table 1). In time studies the input-output ratio of an individual is ob-

served, normally during a rather short period of time. Even when there is no intention to correct data prior to analysis, behavior is intrinsically observed during data collection. However, unless it is included in the experimental design (e.g. experiments are specifically designed to compare methods or analyze the work element distribution), behavior is not normally included in analyses except as circumstantial information or for explaining abnormalities in the data. To enable comparisons of individuals' performance under such settings, there are high requirements for external conditions to be equivalent between time studies, and studies should ideally include repetitions to account for the variation that inevitably occurs, regardless of the care taken to control conditions. Furthermore, including controls and repetitions to exclude as much variation as possible requires substantial time and resources, and time studies are seldom intended to assess individuals (despite the practical potential of such assessments), but rather to predict work performance, often in terms of a productivity norm for a considered cohort of machine operators (cf. standard time for a task). As previously mentioned, the worker's behavior is generally either not considered at all, or it is compensated for in the process of synthesizing work study observations into productivity norms. Irrespective of the methodology used, an established norm enables time studies to be conducted in terms of comparing an individual's observed times with pre-determined norm times for given tasks.

The term performance rating is often used for the time study methodology of establishing productivity norms by subjectively adjusting observed output to account for variations in behavior (see, for instance, Wittering (1973), Bains (1995) or Nieble and Freiwalds (2003) for procedural descriptions). However, here the term refers to the subjective rating of an individual's performance in relation to other individuals. In an assessment an individual is observed by a rater, normally during a rather short period of time, with the main focus on his/her behavior. The rater's assessment should consider external influences, thus decreasing the requirements for equivalent external conditions. Ideally, the rater should also compensate for behavioral variations of an individual. Output is also intrinsically observed during data collection, but is not necessarily a formalized part of the rating (i.e. there is no required measurement of input/output ratios). A limitation of performance rating generally is its strong dependency on the rater's competence and judgment (i.e. the rater's performance), since it has long been known (and well documented) that ratings vary between and within raters (e.g. Barnes 1937; Erler 1985; Arvey and Murphy 1998; Nieble and Freiwalds 2003; Murphy et al. 2004; Roch et al. 2009). However, in work psychology there is optimism regarding the use of subjective rating in assessments of individual performance. The variation in ratings is no longer viewed intrinsically as rater »errors«, but as true variation arising from various sources (Arvey and Murphy 1998). Thus, rating variation can be considered to be a mix of variation in performance by both observed individual and rater. Hence, there is increased recognition that subjective rating does not inevitably introduce rater error or bias, and that rating can often provide valid reflections of individuals' true performance at low cost (Arvey and Murphy 1998).

Gathering data from observations of normal production activities is the core of follow-up studies. The methodology applied can range from self-reporting to use of existent records, with the benefit of requiring little resources for gathering data over long periods of time. Generally, long-term data gathering provides more accurate data about normal performance than short-term studies, because infrequent but expected work components are likely to be included. Moreover, the data gathering should not, ideally, interfere with normal work (i.e. there should be no observer effect) and, thus, minimize the well-known effect that individual performance tends to increase when studied (Mayo 1933; Vöry 1954). However, the level of detail and accuracy of the data acquired are generally lower than when researchers themselves gather data. Computerized automatic data gathering in high-tech forestry machines nowadays offers an attractive alternative for assessing individual performance, since although the data acquired are generally of inferior quality, this is compensated by superior quantity.

Due to the revived acceptance of performance rating, and the emerging potential of automatic gathering of output data, it is of interest to examine the different methodologies of performance assessment in a forest work setting to evaluate their interchangeability, in order to facilitate the selection of appropriate methodology according to the study objectives. Therefore, here we evaluate the correlations between assessments of individual performance by long-term follow-up studies of output and short-term performance rating of behavior. Based on previous research from other fields, our hypotheses for the study are that:

a) the results from the two assessment methods are correlated, and

b) the correlation of performance raters' assessments with follow-up assessments varies between raters.

Table 1 Selected forestry work study methods that allow assessment of individual performance, their use of performance components, observational time duration and level of intentionally included subjective elements

Tablica 1. *Odabrane metode studija rada koje omogućuju procjenu individualne izvedbe, korištene sastavnice, vrijeme trajanja promatranja i razinu hotimično uključenih subjektivnih elemenata*

Method *Metoda*	Performance component – *Sastavnica izvedbe rada*		Duration of observations *Trajanje promatranja*	Subjective elements *Subjektivni elementi*
	Behavior (action) *Zahvat (radnja)*	Output (result) *Output (rezultat)*		
Time study *Studij vremena*	*x*	X	Short – *Kratkoročno*	*x* – X*
Performance rating *Procjena izvedbe*	X	*x*	Short – *Kratkoročno*	X
Historical data (follow up) *Podaci iz prošlosti (dugoročno praćenje)*	–	X	Long – *Dugoročno*	*x*

Note: The main component observed is indicated by an upper case X, the additional component (if observed) by a lower case *x*, and lack of observation of either component by –.
* Depending on whether or not data are subjectively corrected
Bilješka: Glavna promatrana sastavnica označena je velikim slovom X, dodatna sastavnica (ako je promatrana) malim nakošenim slovom x, a nepostojanje promatranja znakom –.
** Ovisno o tome jesu li podaci subjektivno korigirani*

2. Material and Methods – *Materijal i metode*

To address the objective and test the hypotheses, two experts each rated 12 harvester operators' work, then their gradings were compared to the operators' normalized historical output from the two months preceding the rating observations. Here the study setup is only briefly described, since more details are provided in Purfürst (2009).

2.1 Environmental conditions and individuals *Okolišni uvjeti i djelatnici*

All of the data were collected in Germany in 2004 – 2006, during first or second thinnings of pre-marked trees in pine-dominated stands in flat terrain with a cut-to-length system, using similarly-sized harvesters (John Deere 1070, Valmet 901 and Ponsse Beaver) to minimize variation due to differences in machinery. All 12 operators had at least two years of relevant work experience at the time of the study, and most had entered their profession via harvester work education at a vocational school. Their age ranged between 20 – 45 years (median 26 years).

2.2 Observers and grading – *Procjenitelji i ocjenjivanje*

Rater 1 was a 28-year-old male work science researcher and had several years of experience of evaluating harvester work. Rater 2 was a 37-year-old male teacher of harvester work in a vocational school and had six years of experience of training and evaluating harvester operators. Raters assessed harvester operators (i.e. individuals) using the form commonly used in vocational training and professional assessments of harvester operators at the Training Centre for Forest Work and Forestry Technology in Neheim-Hüsten, western Germany. The assessment form included a category of overall performance (i.e. behavior) and 11 subcategories (e.g. harvester head positioning, and speed and carefulness in crane movements). However, in this study only the rating for overall performance was used. For each category, raters graded the operator on a five-level integer scale in which 1 was the best performance and 5 was the worst performance. The scale was constructed so mean performance should correspond to a grade 3. The raters graded each operator having observed him during 2 – 3 hours of work. Both raters observed the operators under their normal working conditions, but from different locations; rater 1 observed from a distance of ca. 25 – 30 m and rater 2 sat in the cab with the operator, as he normally did during his vocational training. The two raters had no knowledge of the operators' prior performance and graded independently of each other without any interactions or calibrations of grades or grading. All operators were visited, observed and graded within a period of one week.

2.3 Historical output data – *Izlazni podaci iz prošlosti*

Output data were collected from normal work through the automatic data recording systems of harvesters. Information on times, dates, harvesting data, operators and software was also collected. The time used for the performance assessment was the productive work time, including interruptions shorter than 15 minutes (PWh_{15}). Data used were stored in defined files according to the StanForD-standard (2007):

*.prd Total harvesting production data.

*.pri Harvesting production data for each individual log and stem.

*.drf Operational monitoring data, covering both work time and repair time data.

*.stm Stem data (length and diameter values)

To enable indications of long-term performance to be obtained, but avoid including performance trends (e.g. long-term performance changes), the work period compared to the raters' grade was set to the 60 days preceding the rating observation. The characteristics of the stands where thinning was carried out by operators are summarized in Table 2. To minimize the influence of differences in environmental factors on the data, observed productivity in the stands was normalized (Eq. 1) relative to the mean expected productivity according to:

$$P_i^N = \left(\frac{P_i}{\hat{P}}\right) \times 100\ (\%) \qquad [1]$$

where P_i is the observed productivity in the stand i (m^3 under bark /PWh_{15}) and the expected productivity in the stand. was calculated as a function of the stand's mean stem size (V, m^3 under bark) according to Eq. 2 (Purfürst 2009).

$$\hat{P} = e^{0.684 \times \ln(V) + 3.543}\ (m^3/PWh_{15}) \qquad [2]$$

Eq. 2 was based on the 12 rated operators' and an additional 20 operators' work (in total 32 operators) in thinnings of pine-dominated forest during a data collection period of three years. The function was based on ca. 65 500 work hours distributed among 3 351 stands with mean stem sizes of 0.04 – 0.32 m^3 (under bark), and explained a significant ($p < 0.001$) proportion of the observed variance ($R^2 = 0.61$).

Since the historical data were time-limited, the productivity in individual stands was time-weighted rather than volume-weighted when calculating the mean performance of an operator. This was done using Eq. 3, in which the observed normalized productivity in a given stand was weighted with the stand proportion of the total work time analyzed for the given operator. In Eq. 3, $\overline{P}^N$ is the time-weighted normalized mean productivity of a given operator and t_i is the time worked in stand i (PWh_{15}). $\overline{P}^N$ and time-weighted absolute mean productivity values during the two-month period are presented in Table 2.

$$\overline{P}^N = \left(\sum_{i=1}^{n} P_i^N \times t_i\right) \times \left(\sum_{i=1}^{n} t_i\right)^{-1}\ (\%) \qquad [3]$$

2.4 Statistical analysis – *Statistička analiza*

Due to the categorical features of the rating scale, non-parametric tests were generally used to avoid violating parametric tests assumptions, e.g. of con-

Table 2 Stand characteristics and time-weighted mean productivity for operators according to data from two months of historical (follow-up) data
Tablica 2. *Sastojinske značajke i vremenski uprosječena proizvodnost za vozače harvestera naspram povijesnih podataka o dugoročnoj proizvodnosti*

Operator *Vozač*	Stands *Sastojine*	Harvested volume *Posječeni obujam*	Work time *Vrijeme rada*	Stem size *Obujam deblovine*	Productivity, mean[a][SD] *Proizvodnost, arit_sred[a][SD]*	
					Absolute *Apsolutna*	Relative *Relativna,* $\overline{P}^N$
	N	m^3	PWh_{15}	mean[a][SD]	m^3/PWh_{15}	%
A	5	2247	203	0.15 (0.02)	11.1 (1.6)	115 (11)
B	11	2118	136	0.22 (0.09)	16.2 (4.6)	132 (10)
C	5	1345	161	0.16 (0.05)	8.3 (1.8)	85 (2)
D	4	1280	195	0.12 (0.07)	7.0 (2.4)	88 (3)
E	8	1383	181	0.15 (0.06)	7.7 (2.1)	84 (13)
F	4	1156	156	0.09 (0.05)	7.2 (2.2)	113 (6)
G	4	1563	132	0.18 (0.13)	11.5 (5.5)	116 (7)
H	5	1318	142	0.17 (0.08)	9.6 (3.1)	96 (4)
I	5	1597	210	0.10 (0.05)	7.5 (2.2)	107 (7)
J	4	1297	162	0.10 (0.05)	8.1 (2.7)	112 (5)
K	3	2132	213	0.14 (0.03)	10.0 (0.8)	114 (7)
L	13	2088	236	0.13 (0.10)	9.0 (4.4)	108 (13)

[a] – time-weighted mean and standard deviation (SD, see Eq. 2)
[a] – *vremenski vagana aritmetička sredina i standardna devijacija (SD, vidi jednadžbu 2)*

tinuous and normally distributed values. Relationships were tested with the Spearman correlation test and illustrated with parametric linear regressions. The significance of differences in ranking between observers was tested with the Wilcoxon signed rank test. In all the non-parametric tests the rank of a given value relative to other values was used instead of the actual value. When assumptions for parametric tests are fulfilled, non-parametric tests can still be used, but they are less good at distinguishing relationships and mean differences. Under such conditions, Wilcoxon and Spearman tests are 5% and 9%, respectively, less efficient than the corresponding parametric tests (paired *T*-test, and Pearson's correlation test) (Zar 1996). SPSS 15.0 (SPSS Inc., U.S.A.) was used for all statistical analyses, with the critical significance level set to 5%.

3. Results – *Rezultati*

There were significant negative relationships ($p < 0.001$) between the long-term relative productivity level and both raters' separate grading of operators (Fig. 1), indicating a congruency between objective long-term measurements and short-term, subjective ratings.

The level of correlation was generally high ($r_s > 0.9$), but slightly different for the two raters, as illustrated by the difference in slopes of the linear regression functions in Fig. 1. Combining the two raters' grading into a mean value for each operator improved the correlation ($r_s = –0.944$, $p < 0.001$) and the combined regression function had a similar slope to the relationship for rater 2. This was mainly due to the off-the-scale grade 0 that rater 2 gave one operator who was considered especially skillful. With that grade transformed to the scale's highest grade for skillfulness (1), the relationship between objective, long-term relative productivity and rater 2's grading remained strong ($r_s = –0.900$, $p < 0.001$), with a slope approaching that of rater 1 ($y = 8.25 – 0.06x$). However, this correction did not result in different correlation values for the mean over raters.

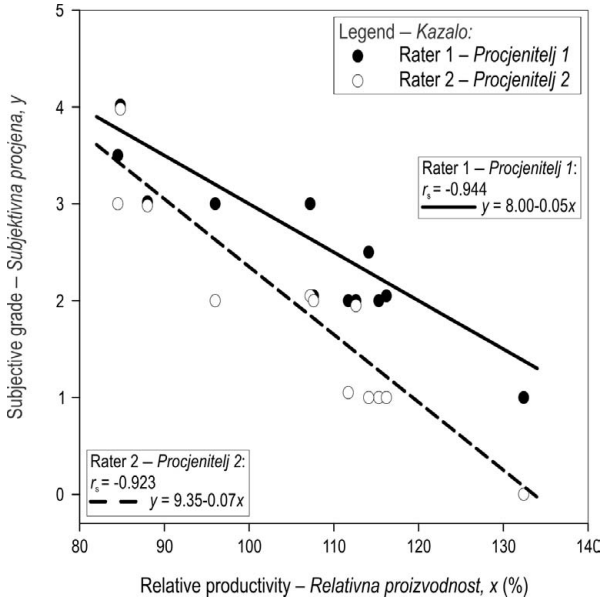

Fig. 1 Relationships between operator performances assessed by normalized long-term, objective productivity data and subjective, short-term rating by the two raters

Slika 1. *Odnos učinkovitosti vozača koju su utvrdila dva procjenitelja normaliziranim dugoročnim praćenjem (objektivne proizvodnosti) i kratkoročnim praćenjem izvedbe*

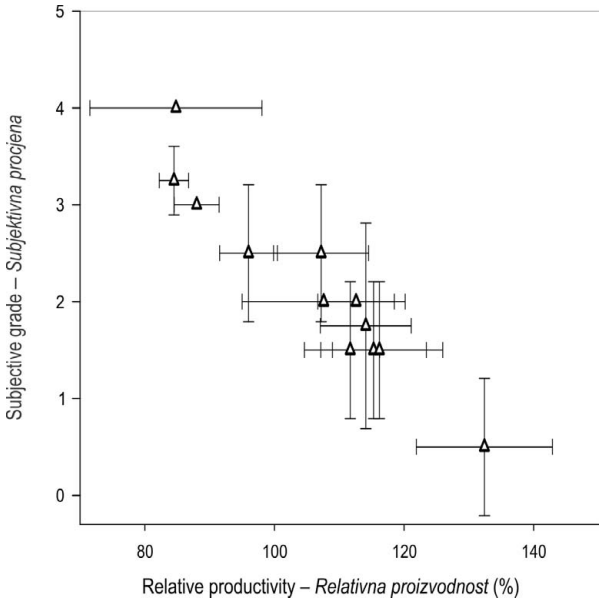

Fig. 2 Relationship between operator's performance assessed by normalized objective, long-term (two months) productivity data and subjective short-term rating (two raters). Triangles indicate mean values for an operator, and bars indicate standard deviations

Slika 2. *Odnos učinkovitosti vozača utvrđene normaliziranim dugoročnim praćenjem (dva mjeseca) podataka proizvodnosti i subjektivnom kratkoročnom procjenom izvedbe (dva procjenitelja). Trokutići označuju prosječne vrijednosti za vozača, a crtane oznake standardne devijacije*

When the variation in the long-term, relative productivity data and the variation in rating due to differences between raters were compared, a clear relationship remained between the objective and subjective assessments (Fig. 2). The variation around the mean relationship seemed, however, to be rather large; the SD-based area was ca. 20% wide and two grades high.

There was a significant positive relationship ($r_s = 0.831$, $p < 0.001$) between the two raters' grading of operators (Fig. 3). However, rater 1 systematically

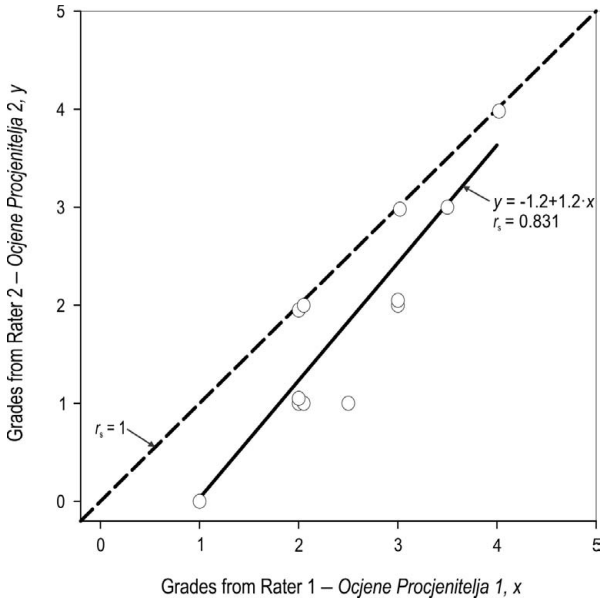

Fig. 3 Relationship between the two raters' grading of harvester operators (N=12; ratings for four operators coincide)

Slika 3. *Odnos ocjena vozača harvestera dvaju procjenitelja (N=12, procjene za četiri operatora koincidiraju)*

gave higher grades than rater 2 (Wilcoxon Signed Rank test, $Z = -2.6$, $p = 0.008$), although the between-grader difference diminished as the grade increased. Based on the regression function in Fig. 3, grades set by rater 2 would have to be increased by (1.2 – 0.2×(grade of rater 2)) to correspond to the mean grading of rater 1.

Both raters identified that the average performance of the operator sample was above average (population average = 100%, observed median = 110%), because their median grades (rater 1 = 2.3, rater 2 = 2) were lower than the scale center (3) (Fig. 4). However, the raters apparently had different reference values for their distribution of grades, as illustrated by the deviation in the distributions of grades between raters in Fig. 4. Grades given by rater 1 can be transformed to long-term normalized productivity intervals because there is only one overlap in productivity level between grades (Fig. 1). Hence, rater 1 seems to have graded according to: 1 > 114%, 2 = 105 – 114%, 3 = 95 – 104%, 4 = 85 – 94% and 5 < 85%. Overlaps were more frequent for rater 2 and, thus, the rater's grades could not be readily transformed to long-term, normalized productivity intervals.

4. Discussion – *Rasprava*

4.1 Results – *Rezultati*

The strong correlation between rater grades and long-term productivity confirms our hypothesis that the results from the two assessment methods are correlated. The results clearly indicate that competent raters can successfully filter the many, interacting behavioral components (e.g. speed and appropriateness

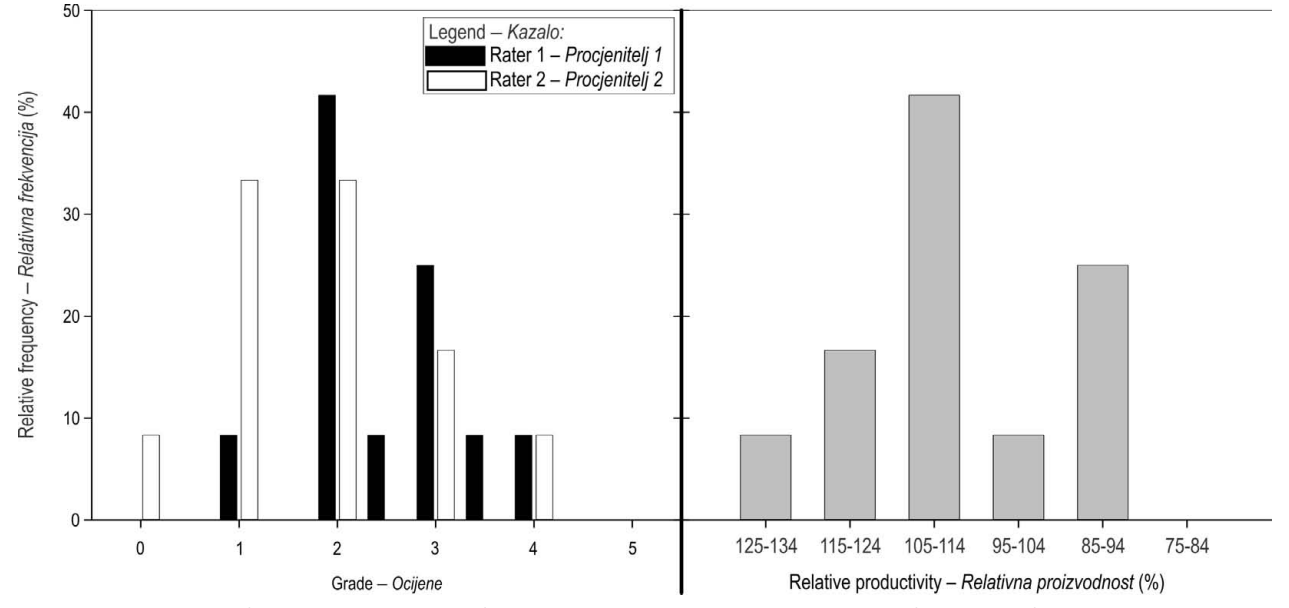

Fig. 4 Relative distributions of the two raters' gradings (left panel) and the relative productivity (right panel) of the sample of operators (N = 12). Expected population mean performance is located at grade 3 and 100% relative productivity, respectively

Slika 4. *Relativne distribucije dvaju procjenitelja (lijevo) i relativna proizvodnost (desno) uzorkovanih vozača (N = 12). Očekivana prosječna vrijednost uzorka smještena je kod ocjene 3 i relativne proizvodnosti od 100 %*

of movements) and translate their rather short-term observations into grades that reflect operators' long--term output. However, our hypothesis that the correlation between the two assessment methods varies between raters was also confirmed. This result is consistent with the vast body of previous research (e.g. Barnes 1937; Murphy et al. 2004; Roch et al. 2009) and shows that despite the high concordance between rater grades, grades are not directly interchangeable. The differences are likely to increase the more graders' perception of normal work performance vary. Moreover, the results demonstrate that the variation in assessed individual performance is rather high, irrespective of the method used. This is consistent with expectations of dynamic performance of individuals (e.g. Arvey and Murphy 1998; Sonnestag and Frese 2002). Hence, for a given operator in this study, the observed variation in long-term productivity is most likely due to a combination of variation in the working conditions that is not accounted for by the normalization standard (Eq. 2, i.e. variation in variables other than mean stem size) and variation in individual performance, whereas the variation in an individual's performance grades is most likely due to performance variation between raters. Hence, there are likely to be two contributory sources of variation in individual performance to the variation in correlation between the methods; one from the operator and one from the rater.

Further, it was interesting to note that both raters clearly deliberately modified their rating scale, according to their perceptions of how to award individuals appropriate grades. Hence, the five-level scale seems to have conflicted with the raters' internal need for detail to obtain appropriate accuracy. This conflict might have been reduced if the grading alternatives had been increased by, for instance, allowing half--grades (e.g. 2.5). However, the need for accuracy should be balanced against the need for simplicity. The scale modifications and rating discrepancies indicate that both the construction and use of scales for subjective performance rating require thorough consideration. The use of scales also implies a need for some kind of rating criteria, which are often quite difficult to formulate because knowledge of the most appropriate behavior under the given conditions is required, and in the complex work of operating a harvester in a highly variable forest environment it is not straightforward to identify, or agree upon, the most appropriate behavior (cf. Ovaskainen et al. 2004). However, the raters obviously drew similar conclusions, although they could have disagreed with each other to some extent if they had provided detailed descriptions of their rating criteria. Such a procedure is consistent with the training and calibration of raters; an obvious thing to do when aiming to ensure that gradings are as similar as possible. Such efforts are likely to decrease the variation within and between raters, but will not eliminate it (e.g. Barnes 1937; Arvey and Murphy 1998).

4.2 Strengths and limitations – *Prednosti i ograničenja*

To our knowledge, this is the first study to address the correlation between subjective, short-term assessments and objective long-term assessments of individuals in forestry work. Compared to many previous studies on harvester work, the number of observed operators (N = 12) was rather high. However, the number of raters should ideally have been higher. Nevertheless, the results provide interesting and valid indications of correlation between the two assessment methods, although the findings, especially the level of correlation, should not be generalized without proper precautions.

From a philosophical standpoint it can be questioned if any methods are truly objective when considering the inherently present subjective elements in, for instance, selecting and applying methods or in the data analyzes. However, the dichotomy of objectivity-subjectivity used in this study is considered to be justified by the differences in intentional subjective influence in the applied methods and is, thus, a relative definition. The influence of subjective features is limited when applying the established long-term methodology, which is therefore considered the most objective method of the two. In contrast, the application of performance rating is constructed to contain an intrinsic element of subjectivity and is therefore considered the most subjective method of the two.

As mentioned in the introduction, a general limitation of the use of historical data is that quantity is obtained at the expense of quality and control. In this study one of the difficulties (which did not affect the results) was that substantial effort was required to handle variation in software and realization of the StanForD standard between harvester manufacturers. Moreover, despite the information and instructions provided to operators about the follow-up study, there was, as expected, little control over how data were recorded and whether or not operators actually managed their harvester computer in the stipulated manner. However, the data recording for the follow-up study did not differ from their normal recording and reporting procedures, so no new and unfamiliar procedures were introduced. Moreover, the productivity levels observed in the total follow-up material (Eq. 2) was reasonably consistent with those reported in previous studies (e.g. Sirén and Aaltio 2003; Nurminen et al. 2006).

4.3 Practical applications – *Praktične primjene*

When considering the practical applications of the study presented here, it is first necessary to recognize the variation in objectives and available resources for work measurements (e.g. Björheden 1991; Nieble and Freiwalds 2003). For instance, Sanders (1975) pointed out that »Work measurements are carried out for a purpose, which may be planning, incentives, controls, costing or some combination of these, and each purpose has its own precision requirement that must be met.« The studied methods assessed individual performance similarly, and either could be used to acquire relevant data for adapting planning, incentives and costing to an individual level in forestry, i.e. either could be used to adjust population norms to the individual's performance, with reasonable accuracy for normal production purposes. Moreover, it could also be possible to both assess the need for vocational training and its effects (for control purposes). This could be done rapidly, with little effort, by performance rating, with the proviso that it would not be automatically possible to compare individuals graded by different raters. The analysis of historical data would probably provide a more accurate estimation of performance over time, but would require greater effort, especially if production data were not normally recorded.

In terms of applications in forest work science, one also has to consider both the objective and the need for accuracy. If the purpose of a study is to rate individual performance, either of the two methods can provide relevant information according to our results. However, if the objective is to minimize the operator effect in order to conduct comparative studies or to construct productivity norms, the appropriateness of the two methods is questionable. If, for instance, it is only possible to compare a machine operated by one individual with another machine operated by another individual, the operator effect has to be handled in order to generalize results. One possible way to do so is to correct output data by dividing acquired values by relative performance levels assessed by either of the two methods. However, the question is how to ascertain whether such a procedure will decrease and not increase the uncontrolled variation in the data. Hence, before designing such a study, one should consider whether the expected uncontrolled variation in operators' performance can be most efficiently minimized by introducing uncontrolled variation in a rater's behavior or in historical data, or if other precautions could be taken to improve the reliability of the data.

5. References – *Literatura*

Anon., 2007: Standard for forest data and communications – StanForD. Skogsforsk. Uppsala, Sweden. 12 p.

Appelroth, S.-E., 1980: Comparability of work study results. In Proceedings from the IUFRO Symposium on Stand Establishment Techniques and Technology in Moscow and Riga 3–8 Sept 1979. IUFRO Group 3.02-00, 414–419.

Appelroth, S.-E., 1989: The analysis and interpretation of forest work study results. In Proceedings of a Symposium on the Equipment/Silviculture Interface in Stand Establishment Research and Operations. Jasper, Alberta. Information Report O-X-40, Ontario Region, Forestry Canada, 173–183.

Arvey, R. D., Murphy, K. R., 1998: Performance evaluation in work settings. Annual Review of Psychology 49(1): 141–168.

Bains, A., 1995: Work measurement – the basic principles revisited. Work Study 44(7): 10–14.

Barnes, R., 1937: Motion and Time Study. 1st ed. John Wiley & Sons, Inc. New York. 285 p.

Beheshti, H. M., Lollar, J. G., 2008: Fuzzy logic and performance evaluation: discussion and application. International Journal of Productivity and Performance Management 57(3): 237–246.

Björheden, R., 1991: Basic time concepts for international comparisons of time study reports. International Journal of Forest Engineering 2(2): 33–39.

Erler, J., 1985: Durchschnittlicher Zeitverbrauch oder Zeitbedarf bei Durchschnittsleistung? Zeitschrift für Arbeitswissenschaft 39(3): 166–168.

Gellerstedt, S., 2002: Operation of the single-grip harvester: motor-sensory and cognitive work. International Journal of Forest Engineering 13(2): 35–47.

Harstela, P., 1975: Factors affecting the consumption of working time and the strain on the worker in some forest work methods. A theoretical and empirical analysis. Communicationes Instituti Forestalis Fenniae 87.2. University of Helsinki, Faculty of Agriculture and Forestry. Helsinki. 130 p.

Kärkkäinen, M., 1975: Foundations of forest work research. A critical review. Research notes No. 31. Department of Logging and utilization of forest products, University of Helsinki. Helsinki. 167 p.

Lindroos, O., 2010: Scrutinizing the theory of comparative time studies with operator as a block effect. International Journal of Forest Engineering 21(1): 20–30.

Mattson Mårn, L., 1953: Arbetsstudier – Ett av arbetslärans viktigaste hjälpmedel [Work studies – One of work science's most important tools]. Skogshögskolan [Royal College of Forestry]. Stockholm. 133 p.

Mayo, E., 1933: The Human Problems of an Industrial Civilization. Macmillan Company. New York.

Murphy, K. R., Cleveland, J. L., Skattebo, A. L., Kinney, T. B., 2004: Raters who pursue different goals give different ratings. Journal of Applied Psychology 89(1): 158–164.

Newman, D. A., Kinney, T. B., Farr, J. L., 2004: Job performance ratings. In Thomas, J. C. (eds) Comprehensive Hand-

book of Psychological Assessment, Vol. 4: Industrial/Organizational Assessment. Wiley. New York. p. 956–1008.

Nieble, B., Freiwalds, A., 2003: Methods, Standards and Work Design. 11th ed. McGraw-Hill. Boston, MA. 747 p.

Nurminen, T., Korpunen, H., Uusitalo, J., 2006: Time consumption analysis of the mechanized cut-to-length harvesting system. Silva Fennica 40(2): 335–363.

Ovaskainen, H., Uusitalo, J., Väätäinen, K., 2004: Characteristics and significance of a harvester operators' working technique in thinning. International Journal of Forest Engineering 15(2): 67–78.

Ovaskainen, H., 2005: Comparison of harvester work in forest and simulator environments. Silva Fennica 39(1): 89–101.

Ovaskainen, H., Heikkilä, M., 2007: Visuospatial cognitive abilities in cut-to-length single-grip timber harvester work. International Journal of Industrial Ergonomics 37(9–10): 771–780.

Purfürst, F. T., 2009: Der Einfluss des Menchen auf die Leistung von Harvestersystemen [The operator's influence on harvester productivity]. PhD-thesis. Institut für Forstnutzung und Forsttechnik, Technische Universität Dresden. Dresden, Germany. 307 p. (In German with English summary).

Reichel, K., 1999: Relative time studies – An empirical survey based on individual tasks during logging operations. Allgemeine Forst Und Jagdzeitung 170(8): 143–148.

Roch, S. G., Paquin, A. R., Littlejohn, T. W., 2009: Do raters agree more on observable items? Human Performance 22(5): 391– 409.

Samset, I., 1990: Some observations on time and performance studies in forestry. Communication 43.5. Norwegian Forest Research Institute. Ås, Norway. 80 p.

Samset, I., 1992: Forest operations as a scientific discipline. Communication 44.12. Norwegian Forest Research Institute. Ås, Norway. 48 p.

Sanders, N. W., 1975: Precision in work measurements. Work Study 8: 15–23.

Sirén, M., Aaltio, H., 2003: Productivity and costs of thinning harvesters and harvester-forwarders. International Journal of Forest Engineering 14(1): 39–48.

Sonnestag, S., Frese, M., 2002: Performance concepts and performance theory. In Sonnestag, S. (eds) Psychological Management of Individual Performance. John Wiley & Sons, Ltd. Chichester, UK. p. 3–25.

Steinlin, H., 1955: Zur Methodik von Feldversuchen im Hauungsbetrieb [Methodology of logging field experiments]. Mitteilungen Band XXXI, heft 2. Schweizerischen Anstalt fur Forstliches Versuchswesen. Zurich. 249–320 p.

Sundberg, U., 1988: The emergence and establishment of forest operations and techniques as a discipline in forest science. Communication 41.8. Norwegian Forest Research Institute. Ås, Norway. 107–137 p.

Tangen, S., 2003: An overview of frequently used performance measures. Work Study 52(7): 347–354.

Tangen, S., 2005: Demystifying productivity and performance. International Journal of Productivity and Performance Management 54(1): 34–46.

Thompson, M. A., 1992: Observation and analysis of performance in forest work. In Proceedings from the IUFRO International Symposium: Work Study – Measurement and Terminology. Göttingen, Germany 10–12 June 1992. Institute of Forest Engineering, Georg-August-University of Göttingen. p. 202–219.

Tubré, T., Arthur, W., Bennett, W., 2006: General models of job performance: Theory and practice. In Bennett, W. and Woehr, D. (eds) Performance Measurements: Current Perspectives and Future Challenges. Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey. p. 175–204.

Wilson, J. R., 1998: A framework and a context for ergonomics methodology. In Wilson, J. R. and Corlett, E. N. (eds) Evaluation of Human Work. Taylor & Francis Ltd. London. p. 1–39.

Wittering, W. O., 1973: Work study in forestry. Forestry Commission Bulletin 47. London, U.K. 100 p.

Vöry, J., 1954: Analysis of the time study materials of some forest jobs. Publication No. 31. The Forest Work Studies Section of the Central Association of Finnish Woodworking Industries, Metsäteho. Helsinki, Finland. 117 p.

Zar, J. H., 1996: Biostatistical Analysis. 3 ed. Prentice-Hall. Upper Saddle River, N.J. 662 p.

Sažetak

Odnos dugoročne proizvodnosti i kratkoročne procjene izvedbe rada vozača harvestera

Radnici su ključni i odlučujući čimbenici provedbe većine proizvodnih sustava pa su njihovi individualni rezultati važni pri procjeni sadašnjih, ali i predloženih sustava šumarskih operacija. Analize radnoga procesa mogu biti korisne za različite namjene, npr. za planiranje, za određivanje stimulacija, za kontrolu ili za određivanje troškova proizvodnje.

Zbog toga su s vremenom razvijene različite metode praćenja rada od kojih sve sadrže elemente i za njihovu primjenu i protiv njihove primjene. To su tri metode: studij vremena, kratkoročna procjena izvedbe rada i metoda

dugoročne proizvodnosti na osnovi vremenskih podataka. Najstarija je metoda metoda studija vremena rada. Metoda praćenja kratkoročne izvedbe subjektivna je metoda na osnovi koje se najčešće donose norme za radove u šumarstvu. Zasniva se na usporedbi različitih individualnih podataka o učinkovitosti i načinima rada. Pri praćenju rada procjenitelj subjektivno ocjenjuje djelatnika s naglaskom na njegove načine rada, i to u kraćem razdoblju. Negativna je strana te metode velika ovisnost rezultata o kvaliteti i iskustvu procjenitelja. Metoda utvrđivanja proizvodnosti na osnovi vremenskih podataka zasniva se na prikupljanju i raščlambi podataka o stvarno ostvarenom radnom procesu. Općenito ta metoda nije zahtjevna kao prve dvije kratkoročne metode te daje realnije podatke.

U ovom se radu uspoređuju zadnje dvije metode praćenja rada 12 vozača harvestera: metoda kratkoročne procjene izvedbe rada i njihova dugoročna proizvodnost (output, obujam posječenoga drva u razdoblju). Istraživanje je provedeno u Njemačkoj od 2004. do 2006. godine. Kratkoročna je procjena rada temeljena na opažanjima svakih nekoliko sati. Procjenitelj broj 1 bio je 28-godišnji znanstvenik iz područja proučavanja rada sa višegodišnjim iskustvom praćenja rada harvestera, dok je procjenitelj 2 bio instruktor za rad harvestera i imao je šestogodišnje iskustvo u obuci vozača. Manual za praćenje imao je ocjenu ukupne radne aktivnosti i još 11 potkategorija ocjene rada vozača. Dugoročna je procjena obavljena prihvatom i raščlambom podataka iz automatiziranoga računalnoga sustava harvestera uz podatke o utjecajnim čimbenicima rada. Vrijeme korišteno u analizi bilo je proizvodno vrijeme rada s prekidima do 15 minuta (PMH_{15}). Podaci su pohranjeni u datotekama u standardu StanForD. Postavljene su dvije hipoteze istraživanja: 1) rezultati su usporedbe dviju metoda procjene povezani i 2) značajno su različite procjene između dvaju procjenitelja u odnosu na dugoročno ostvarenu proizvodnost.

U rezultatima je potvrđena hipoteza o snažnoj povezanosti dviju metoda procjene. Spoznalo se da utjecajni čimbenici mogu filtrirati mnoge interaktivne sastavnice radnoga procesa i prevesti kratkoročna opažanja u ocjene koje odražavaju dugoročnu uspješnost vozača (Spearman's $r_s > 0,9$). Osim toga, utvrđena su značajna odstupanja u ostvarenim rezultatima rada primjenom obiju metoda, koje se najvjerojatnije mogu djelomično pripisati subjektivnoj izvedbi i kod vozača i kod procjenitelja. Dokazano je da obje proučavane metode mogu biti primijenjene za utvrđivanje normi (funkcije proizvodnosti) prema izvedbi pojedinoga vozača, s dovoljnom točnosti za normalnu proizvodnju (planiranje rada). Međutim, u znanstvenom je kontekstu moguće ispitati je li predviđena nekontrolirana varijacija pri radu vozača najučinkovitije minimizirana zbog uvođenja nekontrolirane varijacije u djelovanju procjenitelja i/ili povijesnih (iskustvenih) podataka, ili se moraju poduzeti druge mjere radi poboljšanja pouzdanosti ulaznih podataka.

Ključne riječi: mjerenje učinkovitosti, ocjena operatera, CTL proreda harvesterom, StanForD, vozač harvestera, utjecaj vozača, ljudski čimbenici

Authors' addresses – *Adrese autorâ:*

Thomas Purfürst, PhD.
e-mail: thomas.purfuerst@forst.tu-dresden.de
Institute of Forest Utilization
and Forest Technology
Technische Universität Dresden
D-01737 Tharandt
GERMANY

Assist. Prof. Ola Lindroos, PhD.*
e-mail: ola.lindroos@slu.se
Department of Forest Resource Management,
Swedish University of Agricultural Sciences,
SE-901 83 Umeå
SWEDEN

* Corresponding author – *Glavni autor*

Received (*Primljeno*): September 1, 2010
Accepted (*Prihvaćeno*): January 17, 2011