Available online at www.sciencedirect.com
SCIENCE DIRECT®

International Journal of Industrial Ergonomics 36 (2006) 275–282

International Journal of INDUSTRIAL ERGONOMICS
www.elsevier.com/locate/ergon

Observer reliability of industrial activity analysis based on video recordings

Karolina Kazmierczak[a,b,*], Svend Erik Mathiassen[c], Patrick Neumann[a,b,d], Jørgen Winkel[a]

[a]National Institute for Working Life, 402 72 Gothenburg, Sweden
[b]Department of Design Sciences, Lund Institute of Technology, 221 00 Lund, Sweden
[c]Centre for Musculoskeletal Research, University of Gävle, 90712 Umeå, Sweden
[d]Department of Mechanical and Industrial Engineering, Ryerson University, Toronto, Canada

Received 20 December 2004; received in revised form 12 December 2005; accepted 22 December 2005
Available online 2 February 2006

Abstract

The aim of this study was to assess the agreement between observers analyzing activity patterns during truck engine assembly work based on video recordings. Two observers observed the recordings of nine workers, on the average 2.2 h long, assigning activities to four activity categories. For each activity category data were obtained on the mean duration of uninterrupted sequences of activities and their relative time proportion in the job. This data was analyzed with 2-way crossed ANOVA algorithms to derive the components of variance attributed to disagreement between observers, to differences between filmed subjects, and to residual "unexplained" variance. The latter was interpreted as an estimate of within-observer variability and possible interactions between subject and observer. While the observers disagreed about the overall time proportions for the four activity categories by no more than 3.7% of time, their second-to-second classification disagreed for 13% of the total analysis time. The between-observer variance was small as compared to within-observer variance and the variance between subjects performing the same job. Simulations based on the variance components showed that a group mean of the proportion of direct work could be determined with a standard deviation within 5% of the mean by having two observers analyzing one 2-h video recording once, each.

Relevance to industry

The results of this study may support decision making when designing a reliable video-based analysis of industrial work. Thus, the study helps production engineers, ergonomics practitioners and researchers allocate resources between data collection and data analysis, based on their preferences for precision and power of a particular study.

Keywords: Observer reliability; Activity analysis; Video recordings; Assembly work; Ergonomics

1. Introduction

Methods for collecting, classifying and interpreting data on human performance at work lie at the basis of both ergonomics and engineering (Annett and Stanton, 2000; Mathiassen et al., 2005). "Task analysis" techniques intend to describe and examine individual tasks and activities carried out by human beings within a system (Shryane et al., 2000). Analysis of tasks or activities can provide information on factors affecting human performance as well as the information needs of system designers (Annett and Stanton, 2000).

During recent years manufacturing industry has focused strongly on elimination of losses, including an endeavor to increase the proportion of value-adding (direct) work in the jobs of individuals (Rother and Shook, 2001). However, this may increase work intensity and thus the risk for developing musculoskeletal disorders (Kazmierczak et al.,

*Corresponding author. National Institute for Working Life—West, P.O. Box 8850, SE-402 72 Gothenburg, Sweden. Tel.: +46 31501600; fax: +46 31501610.
E-mail address: karolina.kazmierczak@arbetslivsinstitutet.se (K. Kazmierczak).

0169-8141/$ - see front matter
doi:10.1016/j.ergon.2005.12.006

2005). Reliable assessment of duration of value-adding work is therefore a concern from an engineering as well as ergonomics point of view.

The distribution of activities in the job of an individual can be used as an indicator of 'variation' in the job (Mathiassen and Christmansson, 2004), and as a tool for estimating overall job exposure (Winkel and Mathiassen, 1994; Mathiassen et al., 2005). In the latter case, activity proportions are combined with exposure data for each specific activity. A similar approach can be used to estimate a 'product cycle exposure', that is, the average exposure associated with manufacturing one product (Bao et al., 1996; Mathiassen and Winkel, 1997).

The activity time pattern may be assessed on the basis of workers' own reports, for instance in a diary (Petersson et al., 2000; Balogh et al., 2004; Svendsen et al., 2005) or an interview (Mortimer et al., 1999; Kallio et al., 2000). It can also be determined by expert observations, either on site (Fransson-Hall et al., 1995; Buchholz et al., 1996) or through a video recording that is analyzed afterwards (as in the present study).

Several studies have applied computerized methods for decomposing video recordings into activities (Engström and Medbo, 1997; Christmansson et al., 2002). In general, the purpose has been to assess time proportions of activities without consideration to the exact points in time that a particular activity occurs. However, if the activity decomposition preserves the exact time history of work, it can be synchronized to continuous recordings of measured mechanical exposures (Christmansson et al., 2002; Forsman et al., 2002). This allows determination of exposure patterns for specific activities as well as the expected exposure results when changing proportions of activities in a production.

While a number of studies have addressed aspects of reliability in on-site observations (Van der Beek et al., 1992; Fransson-Hall et al., 1995; Buchholz et al., 1996; de Bruijn et al., 1998) only a few studies have assessed the reliability of activity decomposition from video recordings. Uncertainty in the decomposition is a basic determinant of the credibility of the resulting data, irrespective of whether they concern activity proportions (Medbo, 1998) or estimates of job exposure (Mathiassen et al., 2003). One source of uncertainty resides in different observers arriving at different results when analyzing the same video recording ("between-observer" reliability). Pilot studies suggest that disagreements can be substantial (Medbo, 1998).

Thus, the purpose of the present study was to systematically assess the agreement between observers analyzing truck engine assembly work activity patterns from video recordings.

2. Material and methods

2.1. Subjects and work activities

Video recordings of nine subjects performing the same assembly work at separate workstations were obtained. Each subject was followed for, on average 2.2 h (range 1.3–3.1 h), as part of a larger study on ergonomics and productivity (Neumann et al., submitted). The work consisted of final assembly of truck engines with cycles of about $1\frac{1}{4}$ h each, interspersed by short periods of manual engine transportation on a cart. The equipment and tools used by operators were the same across videos. The operators were filmed on different shifts during different days. Activities in the production were coded based on in situ observations and from video-recordings (Table 1). The activity categories were developed on the basis of a so-called "zero-based" loss analysis (Engström and Medbo, 1997), which aims at distinguishing between value-adding "direct" work, and non-value adding work, such as "indirect" work and disturbances.

2.2. Video analysis system

A previously developed activity analysis system was used to analyze the video recordings ("Videolys" system; Chalmers University of Technology; Engström and Medbo, 1997). The system includes a video camera, a video tape recorder, a TV-monitor and a personal computer. A video recording is analyzed in the computer by an observer who visually identifies all activities and allocates them to categories using software synchronized with the video tape recorder. The registration is done manually by a keystroke

Table 1
Classification of activity categories and activities

Activity categories	Activities	Description
Direct work	Assembly	Principal activities in assembly, e.g. mounting and sub-assembly; getting components and tools at workstation
Indirect work	Getting tools	Any tool acquisition requiring operator to leave workplace
	Getting components	From bench or kit; arranging of components; scrap disposal; plug and unplug at the truck
	Engine transport	To and from the workstation
	Engine adjusting	Raise-lower-rotate engine position
	Quality checking	Visual inspection; making notes; manual checking of the engine
	Quality adjusting	Correction of a problem
	Administrative work	Reading specifications; control of papers; making notes
Disturbances	Material shortage	Miss-picked or missing part/component
	Waiting	Lunch break and other breaks;
	Other	chatting with co-workers
Non work[a]		Periods of technical failures; researcher disturbances incl. subject not in the view

[a]Non-work is included in the analyses since it is a part of the time of the video film.

on buttons in a control window (Engström and Medbo, 1997). The observer marks the end of each activity, which is then also registered as the starting point of the next activity. The system's time resolution is 0.04 s, consistent with a 25 Hz video frame rate.

2.3. Training and analysis

Two observers with an engineering background analyzed the same nine video recordings. One observer had previous experience with a similar analysis system. Their training started with getting acquainted with the software after which they were presented to the activity classification by an 'expert' researcher, and practiced the software for 5 h independently. Afterwards, the two observers watched a representative video sequence together and discussed issues related to the classification of activities. This led to some clarifications and adjustments of the activity descriptions particularly with respect to transition points between activities. Then the two observers practiced the activity analysis independently again using the video recording from one selected work cycle (i.e. more than 1 h of work). Afterwards, the observers discussed their activity analyses, arriving at consensus in cases of disagreement. This entire procedure was intended to correspond to a feasible training procedure in industry.

After the training period, each of the two observers analyzed the video recordings from all nine subjects without being allowed to consult the other. An experienced researcher familiar with the assembly system was available during the entire training and analysis period as a "consultant" who could answer questions and clarify issues in this matter. This person was consulted only on a few occasions during the whole analysis period. The results of the final 'independent' analyses provided the data investigated in the present study. Thus, for each of the nine video recordings, two time histories of activities were available; one for each observer. The analyses were done at the level of activities, which were afterwards merged into activity categories (see Table 1). Fig. 1 presents an example of the time history of activities for one subject according to the two observers. Using these time history files, data were obtained for each activity category on the mean duration of uninterrupted sequences in that category, and the relative time proportion of the activity in the job.

The transition points in time between activity categories were analyzed and summarized for each observer and video recording by means of an Excel macro. The macro generated a file containing information about start and stop times of the four activity categories in the processed video recordings and determined whether the two observers agreed on the activity classification for each single video frame (a time resolution of 0.04 s). For each of the nine video recordings, the time history agreement between observers was summarized in a 4×4 contingency table showing the opinion of observer A by activity category (columns) versus that of observer B (rows).

Fig. 1. An example of the activity time history for a period of 10 min in one subject according to the two observers.

2.4. Statistical analysis

For each parameter and activity category, the results from the two observers were entered in a 2-way crossed random-effects model (observer $\times$ subject) to estimate the variance caused by systematic disagreement between observers, the variance due to differences between filmed subjects, and the residual variance. These variance components were determined using ANOVA algorithms. The residual, 'unexplained', variance included within-observer variability and possible interaction between subject and observer, that is relationships between observers that depend systematically on the analyzed subject. The ANOVA algorithms can lead to negative estimates of variance components, and in those cases the variance was set to zero (cf. Searle et al., 1992). Standard deviations (SD) and coefficient of variations (CV) were calculated on the basis of the variance components.

3. Results

Table 2 shows the relative time proportion of the four activity categories. Table 3 shows the mean duration of uninterrupted sequences in each category, for the nine subjects according to the two observers.

Table 4 shows the mean value across observers and subjects for both parameters by activity category, as well as the between-observers, between-subjects, and residual variabilities, expressed as variance, standard deviation (SD), and coefficient of variation (CV%, i.e. SD divided by the mean and multiplied by 100).

Table 5 illustrates the time history agreement between the two observers. In total, the observers agreed on the activity category for 7055.4 s of a total of 8087 s of the video recordings (i.e. 87% agreement), thus disagreeing for 13% of the total time. As an example, 67.5% of the total analysis time was agreed to be direct work by both observers. They both also agreed that 23.7% of total time should be classified in other activity categories than direct work. For 8.6% of total time, one of the

Table 2
Activity proportions (percent observation time), by subject ($n = 9$) and observer (A and B)

Subject	Direct work		Indirect work		Disturbances		Non-work	
	A	B	A	B	A	B	A	B
1	65.0	68.0	21.2	20.4	5.4	2.3	8.5	10.7
2	68.9	66.1	14.6	15.3	11.3	3.4	5.2	15.2
3	71.4	73.2	17.2	17.0	4.7	1.0	6.7	8.9
4	72.4	74.4	20.4	17.7	2.4	2.9	4.7	5.0
5	79.9	74.0	14.3	17.8	2.5	0.0	3.3	8.2
6	70.5	70.3	20.4	18.8	4.5	3.7	4.7	7.4
7	77.8	71.7	13.5	16.0	1.5	1.1	7.1	11.3
8	75.7	70.0	18.2	16.0	2.1	2.5	4.1	11.4
9	75.3	71.0	17.3	20.6	3.6	2.1	4.0	6.2

Table 3
Mean duration (s) of uninterrupted sequences for each activity, by subject ($n = 9$) and observer (A and B)

Subject	Direct work		Indirect work		Disturbances		Non-work	
	A	B	A	B	A	B	A	B
1	41.3	59.5	57.2	31.7	73.0	11.2	36.8	40.0
2	50.7	55.0	63.7	90.3	383.8	70.5	69.7	72.5
3	60.1	75.2	134.6	88.4	267.2	21.2	54.2	38.0
4	47.7	74.8	100.6	86.4	54.4	60.4	50.1	36.8
5	85.2	74.1	56.8	60.2	72.0	0.0	46.8	46.3
6	38.8	56.2	83.0	92.8	91.4	89.7	46.2	30.0
7	73.6	65.4	81.4	104.7	23.2	26.1	37.4	34.7
8	45.1	58.9	58.0	48.7	75.6	58.1	33.3	41.8
9	49.9	57.5	110.1	95.5	273.1	13.9	22.6	29.6

observers classified what he saw as 'direct work' while the other did not.

4. Discussion

This study showed, in general, a good agreement between observers, both on overall activity proportions and on the mean duration of sequences in most activity categories, following the present standardized training period of the observers. The variance between filmed subjects was larger than that between observers in most combinations of parameter and activity category. The residual variance, which we interpret as mainly being due to within-observer (test–retest) variability was, in general, larger than the between-observer variability.

While the present study addressed observer agreement on activity categories in an engineering context, the issue is also relevant to an ergonomist. Thus, several studies have assessed 'task exposures' for the purpose of identifying targets for ergonomics intervention (Dempsey and Mathiassen, 2006) or estimating exposures in combinations of tasks (Mathiassen et al., 2005; Svendsen et al., 2005). These approaches require that job activities are identified, for instance by video observation (Forsman et al., 2002), and hence reliable procedures for that purpose. However, in the context of ergonomics epidemiology focusing explicitly on hazardous exposures, activities would be classified according to, e.g. the occurrence of awkward postures (Van der Beek and Frings-Dresen, 1998; Svendsen et al., 2005), rather than according to their value-adding properties.

4.1. Methodological considerations

The between-observer variance was small, which implies that across a large number of analyses, different observers will reach mean values that do not differ a lot. But, as shown by the large residual variance, observers may disagree substantially on the results for a particular video recording.

This residual, "unexplained" variance was interpreted as a measure of within-observer variability, based on the assumption of a marginal systematic interaction between observers and subjects. A hypothetical example of such interaction could be that a particular observer believes that Swedish workers, as a rule, are lazy, and thus records them as spending more time in indirect work.

The general tendency that within-observer variability was larger than between-observer variability stands in contrast to results of a study of video-based activity analysis in car parts assembly by Medbo (1998). Medbo's study suggested that the variance within an observer when making repeated analyses is generally lower than the variance between observers for activity proportions, frequencies and mean duration of sequences. Medbo reported ratios of within- to between-observer variance between 0.01 and 28.0 for these parameters in 19 different activity categories. Another study by Chaikumarn (2001) also indicated a good consistency when one observer performs repeated activity analyses on the same video recording. The studies of Medbo (1998) and Chaikumarn (2001) explicitly addressed within-observer variance, while we estimated within-observer variability indirectly by the residual variance in our statistical model. Furthermore, in our study the analysis was made on a long-cycle work (h), while the mentioned studies analyzed short-cycle work (min). As discussed above this may have led to some overestimation of this source of variability.

Differences between observers can be caused by different understandings of the activity definitions. Thus, in our study, one observer had been involved in data collection while the other had not. This source of disagreement is probably influenced by the number and complexity of activities; as complexity increases, disagreement between observers can be expected to increase (Kilbom, 1994; Winkel et al., 1995). The motivation for making the analyses may also vary between observers, which can influence for instance the speed of analyses, and the

Table 4
Mean and variance components for the two investigated activity parameters (proportion and mean duration of sequences) and the four activity categories; VAR, variance, SD, standard deviation, CV, coefficient of variation (percent)

	Direct work	Indirect work	Non-work	Disturbances
Proportion (%)				
Mean	72.0	17.6	7.4	3.2
Between-observer variability				
VARbo	1.3	0.0[a]	7.5	1.8
SDbo	1.1	0.0[a]	2.7	1.4
CVbo	1.6	0.0[a]	37.2	42.9
Between-subject variability				
VARbs	7.7	3.1	1.8	1.7
SDbs	2.8	1.8	1.3	1.3
CVbs	3.9	10.1	18.2	40.9
Residual variability				
VARr	6.9	2.8	4.6	3.5
SDr	2.6	1.7	2.1	1.9
CVr	3.6	9.5	29.0	58.7
Mean duration of sequence(s)				
Mean	59.4	80.2	42.6	92.5
Between-observer variability				
VARbo	35.1	0.0[a]	0.0[a]	4800.9
SDbo	5.9	0.0[a]	0.0[a]	69.3
CVbo	10.0	0.0[a]	0.0[a]	74.9
Between-subject variability				
VARbs	77.4	409.4	128.9	320.4
SDbs	8.8	20.2	11.4	17.9
CVbs	14.8	25.2	26.7	19.4
Residual variability				
VARr	79.3	274.8	47.5	8267.7
SDr	8.9	16.6	6.9	90.9
CVr	15.0	20.7	16.2	98.3

[a]The ANOVA gave negative estimates thus the variance set to zero.

Table 5
Time history agreement between the two observers (seconds; in parentheses: percent of total analysis time). Each cell contains the average of the results from the individual tables ($n = 9$). In italics: percent time that the two observers agreed

Observer A Observer B	Direct work	Indirect work	Disturbances	Non-work	Sum observer A
Direct work	5461.5 (*67.5%*)	250.9 (3.1%)	39.0 (0.5%)	2.6 (0.03%)	5754.0 (71.1%)
Indirect work	301.0 (3.7%)	1027.2 (*12.7%*)	59.4 (0.7%)	11.0 (0.1%)	1398.6 (17.3%)
Disturbances	9.8 (0.1%)	4.1 (0.05%)	140.1 (*1.7%*)	19.2 (0.2%)	173.2 (2.1%)
Non-work	97.9 (1.2%)	100.1 (1.2%)	136.6 (1.7%)	426.7 (*5.3%*)	761.3 (9.4%)
Sum observer B	5870.2 (72.6%)	1382.3 (17.1%)	375.1 (4.6%)	459.5 (5.7%)	

willingness to backtrack the tapes to reassess difficult parts. In our study, one observer intended to use the data in her own future research while the other was a temporary employee at the department.

Although complete agreement is strived for, it is obviously impossible to reach. In our study the time history agreement between observers was 87%. This is in the same order of magnitude as the 80% agreement between observers reported by Buchholz et al. (1996) in a study of postures in construction work.

Variability between observers could probably be reduced with an even more careful and exact definition of activities. Thus, Burt and Punnett (1999) suggested that between-observer reliability of postural observations in manufacturing jobs can be minimized when operational definitions are simple and unambiguous; when longer and multiple training sessions precede data collection; and when the number of observed (in their case) postures and the level of detail is limited. Other studies have shared these viewpoints suggesting an extensive common training for observers in order to reduce between-observer variability, and a pilot investigation of reliability (Van der Beek et al., 1992; Medbo, 1998). In our study, the between-observer variability was low; thus it seems that the training procedure was effective. Since we had access to only two observers, however, further studies would be needed to corroborate our estimates of between-observer variability.

The present study suggested that within-observer variance can be substantial. Tang (2000) stated that the magnitude of performance variance might be due to uncontrolled factors in training protocols such as individual differences in experiences or intelligence, and random errors. Ways to reduce within-observer variability could thus be better training of observers, including instructions for performing the analysis more carefully, for instance taking more breaks or taking time to make double-checks in case of uncertainty. A digital video interface may also speed up error correction and foster improved precision.

4.2. Applications

Variance components in measures, such as those derived in the present study, allow for the good practice of analyzing different design options before implementing a research study (Mathiassen et al., 2002). As an example, we made a simulation to assess the trade off between including more observers performing activity analysis in a study, or more filmed subjects (Fig. 2). This example illustrates the case that a pool of observers is available for data analysis. Each video recording is then analyzed one or more times by one or more observers selected at random from this pool. In that case, the variance of the group mean is given by

$$\mathrm{var_m} = (\mathrm{var_{bs}} + \mathrm{var_{bo(gross)}})/n_\mathrm{s}, \tag{1}$$

where $\mathrm{var_{bs}}$ is the variance between subjects (films), $\mathrm{var_{bo(gross)}}$ the gross variance between observers; cf. Eq. (2) below, n_s the number of films (subjects).

The gross variance between observers was estimated as

$$\mathrm{var_{bo(gross)}} = (\mathrm{var_{bo}} + \mathrm{var_{wo}}/n_\mathrm{a})/n_\mathrm{o}, \tag{2}$$

where $\mathrm{var_{bo}}$ is the variance between observers, $\mathrm{var_{wo}}$ the variance within observers, n_a the number of times each observer repeats a particular analysis, n_o the number of observers analyzing each video recording.

These algorithms were used on our data concerning proportion of direct work (group mean 72%, cf. Table 4). As seen in Fig. 2, a SD-value corresponding to 3.6% time (that is, 5% of the group mean), could be obtained by having two observers both analyze the recording from a single subject. This level of precision might satisfy a company screening the performance of its production system. Thus the company would have to collect video films for two cycles (about 2 h) for one subject. Since 1 h of video recording requires about 3 h of analysis by an experienced observer (Engström and Medbo, 1997), this would imply a total of about 12 h of analysis time for two observers, each one being engaged for 6 h.

A researcher would probably be interested in a better precision, for instance an SD corresponding to two percent of the mean, i.e. 1.4% time. This level could be reached, either by filming six subjects and have them analyzed by two observers, or by filming eight subjects and let one observer analyze each film. The two alternatives require about 12 and 16 h of data collection respectively and in total about 72 and 48 h of analysis time, respectively.

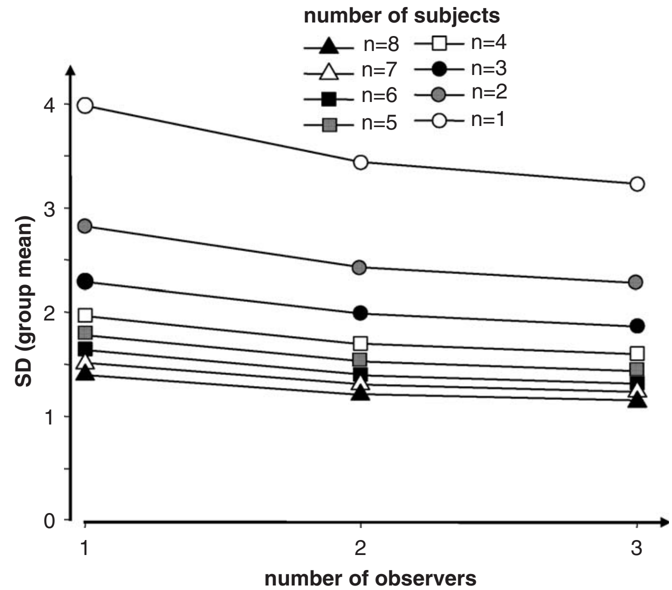

Fig. 2. Estimated standard deviation (SD) of the group mean proportion of direct work according to the number of filmed subjects (family of curves) and the number of observers analyzing the films (abscissa). Estimates are based on the variance components in Table 4, using Eqs. (1) and (2). Each observer analyzes a particular film once.

As another practical illustration, we assessed the necessary measurement resources needed to obtain a given sensitivity of a study investigating the difference between two independent groups. Power analyses similar to this are commonly used tools for decisions on study design at the planning stage. In our hypothetical case, a company plans to rationalize production in order to increase the mean proportion of direct work from the present 72–77% of the total time. Our calculations, based on standard power statistics (Mathiassen et al., 2002), showed that in order to detect the difference of 5% time at a power $(1-\beta)$ of 0.8 and a p-level α of 0.05, 22 subjects would be needed (11 "before" and 11 "after" the rationalization) if the films were analyzed once each by one observer. The same sensitivity to change could be achieved, if in total, 18 films (9 "before" and 9 "after") were analyzed twice by one observer, or if 16 films (8 "before" and 8 "after") were analyzed once by two observers. Quantitative comparisons such as these can be used further for considering which option might be more feasible, for instance in terms of time, costs, and accessibility to subjects and trained observers. The above case applies to a company which would accept to make a large number of analyses in order to secure a sufficient power level. However, a smaller number of analyses, i.e. 2 films "before" and 2 "after" would be "good enough" if the company accepts a considerable risk of not being able to detect the intervention effect even if it exists, or erroneously claim an effect that is not true.

4.2.1. Time history agreement

The analyses of agreement between the two observers on the time history of activity categories showed notable differences for direct and indirect work (Table 5). Apparently, observers may agree reasonably well on total proportions while they disagree on the occurrence in time of a particular activity category. For instance, the two observers arrived at direct work totals of 71.1% time and 72.6% time, respectively, across all nine subjects, a difference of 1.5% time (Table 5). However, for 8.6% of the total analysis time, the observers disagreed whether the activity performed at a particular point in time classified as direct work or not.

Time history agreement is important when the activity analysis is synchronized to other data sources, for instance in order to determine activity exposures from recordings of physical workload (Winkel et al., 1999; Forsman et al., 2002; Kazmierczak et al., 2005). In this case, disagreement on the exact times of transitions between activities leads to misclassification of activities, and thus to flawed activity exposures.

5. Conclusions

The study shows the level of agreement between observers that can be expected when using the investigated training procedure for analyzing video recordings with respect to value-adding time in long-cycle assembly work. The between-observer variance was small compared to both within-observer variance and gross variance between subjects.

This information is useful when allocating resources for data collection and analysis in general screenings of production system performance, and in ergonomics intervention research.

Acknowledgments

This work was financially supported by National Institute for Working Life in Sweden. The authors would like to thank Volvo Powertrain Corp. who made the data collection possible. The authors would like to thank Dr. Mikael Forsman for his help with developing data analysis macros. Thanks also to Montakarn Chaikumarn and Olof Persson for performing the activity analyses.

References

Annett, J., Stanton, N.A., 2000. Task Analysis. Taylor & Francis, London.

Balogh, I., Ørbæk, P., Ohlsson, K., Nordander, C., Unge, J., Winkel, J., et al., 2004. Self-assessed and directly measured occupational physical activities-influence of musculoskeletal complaints, age and gender. Applied Ergonomics 35, 49–56.

Bao, S., Mathiassen, S.E., Winkel, J., 1996. Ergonomic effects of a management-based rationalization in assembly work—a case study. Applied Ergonomics 27, 89–99.

Buchholz, B., Paquet, V., Punnett, L., Lee, D., Moir, S., 1996. PATH: a work sampling-based approach to ergonomic job analysis for construction and other non-repetitive work. Applied Ergonomics 27, 177–187.

Burt, S., Punnett, L., 1999. Evaluation of interrater reliability for posture observations in a field study. Applied Ergonomics 30, 121–135.

Chaikumarn, M., 2001. Variability in activities and postures during electronic assembly work assessed by video-based analysis system. Master Thesis 109 CIV, Luleå University of Technology.

Christmansson, M., Medbo, L., Hansson, G-Å., Ohlsson, K., Unge Byström, J., Möller, T., et al., 2002. A case study of a principally new way of materials kitting—an evaluation of time consumption and physical workload. International Journal of Industrial Ergonomics 30, 49–65.

De Bruijn, I., Engels, J.A., van der Gulden, J.W.J., 1998. A simple method to evaluate the reliability of OWAS observations. Applied Ergonomics 29, 281–283.

Dempsey, P.G., Mathiassen, S.E., 2006. On the evolution of task-based analysis of manual materials handling, and its applicability in contemporary ergonomics. Applied Ergonomics 37, 33–43.

Engström, T., Medbo, L., 1997. Empirical analysis and assessment of automotive assembly work: an analysis procedure for work efficiency and ergonomic aspects. In: Seppälä, P., Luopajärvi, T., Nygård, C.-H., Mattila, M. (Eds.), International Ergonomics Association; Proceedings of the 13th Triennial Congress—IEA 97, Organizational Design and Management, Tampere, Finland, pp. 171–173.

Engström, T., Medbo, P., 1997. Data collection and analysis of manual work using video recording and personal computer techniques. International Journal of Industrial Ergonomics 19, 291–298.

Forsman, M., Hansson, G-Å., Medbo, L., Asterland, P., Engström, T., 2002. A method for evaluation of manual work using synchronised video recordings and physiological measurements. Applied Ergonomics 33, 533–540.

Fransson-Hall, C., Gloria, R., Kilbom, Å., Winkel, J., Karlqvist, L., Wiktorin, C., Stockholm Music 1 Study Group, 1995. A portable ergonomic observation method (PEO) for computerized on-line recording of postures and manual handling. Applied Ergonomics 26, 93–100.

Kallio, M., Viikari-Juntura, E., Häkkänen, M., Takala, E.P., 2000. Assessment of duration and frequency of work tasks by telephone interview: reproducibility and validity. Ergonomics 43, 610–621.

Kazmierczak, K., Mathiassen, S.E., Forsman, M., Winkel, J., 2005. An integrated analysis of ergonomics and time consumption in Swedish 'craft-type' car disassembly. Applied Ergonomics 36, 263–273.

Kilbom, Å., 1994. Assessment of physical exposure in relation to work-related musculoskeletal disorders—what information can be obtained from systematic observation? Scandinavian Journal of Work Environment and Health 20, 30–45.

Mathiassen, S.E., Winkel, J., 1997. Ergonomic exposure assessment adapted to production system design. In: Seppälä, P., Luopajärvi, T., Nygård, C.-H., Mattila, M. (Eds.), International Ergonomics Association; Proceedings of the 13th Triennial Congress—IEA 97, Organizational Design and Management, Tampere, Finland, pp. 195–197.

Mathiassen, S.E., Burdorf, A., Van der Beek, A.J., 2002. Statistical power and measurement allocation in ergonomic intervention studies assessing upper trapezius EMG amplitude, a case study of assembly work. Journal of Electromyography and Kinesiology 12, 45–57.

Mathiassen, S.E., Burdorf, A., Van der Beek, A.J., Hansson, G-Å., 2003. Efficient one-day sampling of mechanical job exposure data—a study based on upper trapezius activity in cleaners and office workers. American Industrial Hygiene Association Journal 64, 196–211.

Mathiassen, S.E., Christmansson, M., 2004. Variation and autonomy. In: Delleman, N., Haslegrave, C., Chaffin, D. (Eds.), Working Postures and Movements—Tools for Evaluation and Engineering. Taylor & Francis, London, pp. 330–366 (Chapter 11.2).

Mathiassen, S.E., Nordander, C., Svendsen, S.W., Wellman, H.M., Dempsey, P.G., 2005. Estimating mechanical job exposure: a task-based model compared to the occupational group mean. Scandinavian Journal of Work. Environment Health 31, 138–151.

Medbo, P., 1998. Reliabilitet vid videoanalys, materialflödesanalyser med hjälp av videoinspelningar [in Swedish; Reliability of video analysis, material flow analyses with help of video recordings]. Report. Chalmers University of Technology, Sweden.

Mortimer, M., Wigaeus Hjelm, E., Wiktorin, C., Pernold, G., Kilbom, Å., Vingård, E., et al., 1999. Validity of self-reported duration of work postures obtained by interview. Applied Ergonomics 30, 477–486.

Neumann, W.P., Winkel, J., Medbo, L., Mathiassen, S.E., Magneberg, R. Design elements influencing productivity and ergonomics in a case study of parallel and serial flow production strategies, submitted for publication.

Petersson, N.F., Mathiassen, S.E., Björing, G., Winkel, J., 2000. The accuracy of self-rating of exposure to repetitive work. International Journal of Industrial Ergonomics 25, 239–246.

Rother, M., Shook, J., 2001. Lära sig se: att kartlägga och förbättra värdeflöden för att skapa mervärden och eliminera slöseri: en handbok för praktisk tillämpning av metoder och verktyg för Lean produktion. Lean Enterprise Institute Sweden, Göteborg.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Shryane, N.M., Westerman, S.J., Crawshaw, C.M., Hockey, G.R.J., Sauer, J., 2000. Task analysis for the investigation of human error in safety-critical software design: a convergent methods approach. In: Annett, J., Stanton, N.A. (Eds.), Task Analysis. Taylor & Francis, London, pp. 191–208 (Chapter 11).

Svendsen, S.W., Mathiassen, S.E., Bonde, J.P., 2005. Task-based exposure assessment in ergonomic epidemiology—a study of upper arm elevation in the jobs of machinists, car mechanics, and house painters. Occupational and Environment Medicine 62, 18–27.

Tang, K-H., 2000. Computer-based group training and prior achievement: effects on performance means and variances. International Journal of Industrial Ergonomics 26, 467–476.

Van der Beek, A.J., van Gaalen, L.C., Frings-Dresen, M.H.W., 1992. Working postures and activities of lorry drivers: a reliability study of on-site observation and recording on a pocket computer. Applied Ergonomics 23, 331–336.

Van der Beek, A.J., Frings-Dresen, M.H.W., 1998. Assessment of mechanical exposure in ergonomic epidemiology. Occupational and Environmental Medicine 55, 291–299.

Winkel, J., Mathiassen, S.E., 1994. Assessment of physical work load in epidemiologic studies: concepts, issues and operational considerations. Ergonomics 37, 979–988.

Winkel, J., Balogh, I., Hansson, G.-Å., Asterland, P., Kellerman, M., Byström, J., Ohlsson, K., 1995. Assessment of mechanical exposures in ergonomic epidemiology: a research program and some preliminary data. In: Grieco, A., Molteni, G., Piccoli, B., Occhipinti, E. (Eds.), Work with Display Units 94, Fourth International Scientific Conference on Work with Display Units, Milan, Italy, pp. 167–172.

Winkel, J., Christmansson, M., Cyren, H., Engström, T., Forsman, M., Hansson, G-Å., et al., 1999. A Swedish Industrial Research Program 'Co-operative for Optimization of Industrial Production Systems Regarding Productivity and Ergonomics' (COPE). American Journal of Industrial Medicine Supplement 1, 82–85.