

Inter- and intra- observer reliability of risk assessment of repetitive work without an explicit method



Kristina Eliasson ^{a,*}, Peter Palm ^b, Teresia Nyman ^{a,b}, Mikael Forsman ^{c,d}

^a School of Technology and Health, KTH Royal Institute of Technology, Huddinge, Sweden

^b Department of Medical Sciences, Occupational and Environmental Medicine, Uppsala University and Uppsala University Hospital, Sweden

^c IMM Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

^d Centre for Occupational and Environmental Medicine, Stockholm County Council, Sweden

ARTICLE INFO

Article history:

Received 30 March 2016
Received in revised form
30 January 2017
Accepted 2 February 2017
Available online 16 February 2017

Keywords:

Inter-observer reliability
Intra-observer reliability
Risk assessment
Observational methods

ABSTRACT

A common way to conduct practical risk assessments is to observe a job and report the observed long term risks for musculoskeletal disorders. The aim of this study was to evaluate the inter- and intra-observer reliability of ergonomists' risk assessments without the support of an explicit risk assessment method. Twenty-one experienced ergonomists assessed the risk level (low, moderate, high risk) of eight upper body regions, as well as the global risk of 10 video recorded work tasks. Intra-observer reliability was assessed by having nine of the ergonomists repeat the procedure at least three weeks after the first assessment. The ergonomists made their risk assessment based on his/her experience and knowledge. The statistical parameters of reliability included agreement in %, kappa, linearly weighted kappa, intraclass correlation and Kendall's coefficient of concordance. The average inter-observer agreement of the global risk was 53% and the corresponding weighted kappa (K_w) was 0.32, indicating fair reliability. The intra-observer agreement was 61% and 0.41 (K_w). This study indicates that risk assessments of the upper body, without the use of an explicit observational method, have non-acceptable reliability. It is therefore recommended to use systematic risk assessment methods to a higher degree.

© 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Deficiencies in the work environment contribute to the development of musculoskeletal disorders (MSDs), which can have economic consequences for the individual, society and employers. Work-related exposures such as repetitive work, forceful exertions, awkward postures, and vibration, as well as psychosocial and organisational factors are related to the development of MSD (Bongers et al., 2006; Bovenzi, 2006; Lang et al., 2012; Palmer and Smedley, 2007; Punnett and Wegman, 2004; Putz-Anderson et al., 1997; van Rijn et al., 2009a, b; 2010). According to safety and health legislation and recommendations, regular risk assessments should be carried out to identify and prevent potentially harmful work tasks (European Council, 1989). Risk assessment is also an important tool when planning and prioritising work environment interventions such as changes in the physical design of the workplace, in work technique or in

work organisation. Sometimes these interventions can lead to an extensive investment for the employer. After interventions, new risk assessments may be carried out for evaluation purposes. Furthermore, work environment authorities also uses risk assessments when legislative measures are taken towards an employer. It is therefore highly important that risk assessments are valid and reliable.

Ergonomists employed in occupational health services (OHS) often perform risk assessments of physical work environments. Observational methods are described as useful for identifying and assessing potentially harmful occupational exposures due to their low cost and ability to present the result in a way that is easy to understand (e.g. in different risk levels). Several observational methods have been developed for the identification and quantification of physical exposures at work (Dempsey et al., 2005; Neumann, 2007; Takala et al., 2010). Inter-observer reliability studies of different observational methods show mixed results and comparisons between the studies are hampered by differences in the choice of statistical methods. (David et al., 2008; Comper et al., 2012; Spielholz et al., 2008; Stevens et al., 2004; Paulsen et al., 2015).

* Corresponding author.

E-mail address: kristie@kth.se (K. Eliasson).

Very few observational methods have been evaluated with regards to their predictive validity, i.e. do more adverse risk scores predict increased incidence of MSD (Takala et al., 2010). Nevertheless, most observation methods purport to include the dimensions of amplitude, frequency and duration of harmful exposure, and assume that the higher combined score, the higher the risk of MSD.

Different observational methods assess different types of exposure (manual handling, repetitive work etc.) and the selection or combination of methods should be based on the need of the assessment and the exposure type (Takala et al., 2010). In several observational methods different exposure parameters are observed and rated, and then those parameters are used to calculate a total score which is then converted to different risk levels, e.g. green, yellow or red.

As for usage, a web-survey among Swedish ergonomists in 2012 revealed that knowledge about and use of different risk assessment methods was relatively low (Eliasson et al., unpublished manuscript). The study further indicated that ergonomists often assess risks in the work environment solely by means of observation, based on his/her knowledge and experience, without the use of any systematic methodology or explicit method. The results of that study are in agreement with other studies (Wells et al., 2013; Whysall et al., 2004). Furthermore, Whysall et al. (2004) reported that evaluation of implemented recommendations is rare. When risk assessments are reported back to the client is it often in the manner of risk levels defined by a “traffic light” scale, where red = high risk/immediate action is needed; yellow = medium risk/investigate further; or green = low risk/acceptable exposure, which are the levels proposed in the Ergonomic Provisions from the SWEA (Hägg, 2003; Koningsveld et al., 2005; Lind and Rose, 2016; SWEA, 2012). However, in contrast to when systematic observational methods are used, these risk levels are empirically derived, and not based on a calculated score from ratings of different exposure parameters.

One important risk factor for MSD, especially in the neck and upper extremities, is exposure to repetitive work (Nordander et al., 2013; Palmer et al., 2007; Palmer and Smedley, 2007; van Rijn et al., 2009a, b; 2010). However, movements occurring in repetitive work, for example movement velocity, are more difficult to assess using observation compared to assessments of exposures that include macro-postures (Ketola et al., 2001; Lowe, 2004; Spielholz et al., 2001; Takala et al., 2010). Consequently, seeing that risk assessments of repetitive work can be difficult to perform and that assessments are often made without the use of an explicit method, it is of interest to analyse how ergonomists' own “expert”

assessments of repetitive work conform between different ergonomists and different assessment occasions.

The overall aim of the present study was to investigate the inter-observer and intra-observer reliability of risk assessments performed by ergonomists without the use of an explicit observational method.

2. Method

2.1. Observers

In total, 21 OHS-ergonomists participated as observers in the present study. They were recruited through contact with different OHS companies and through social media posts to members of the Swedish Ergonomist and Human Factors Society (EHSS). Employment at an OHS (or equivalent) and at least one year of work experience in the sector, including experience with risk assessments, were the necessary requirements for observer participation in the study. Details about the observers are presented in Table 1.

2.2. Video recorded work tasks

Ten different work tasks from various job sectors were selected (i.e. grocery store shop assistant, meat cutting, industrial assembly, cleaning, post sorting and hairdressing; Table 2). The work postures and movements were mainly of a repetitive character.

Each work task was recorded using two to four video cameras from different angles to enable the best possible conditions for the risk assessments. For each work task, the different views were synchronised into one video with multiple frames to show the different views of the worker with a close-up on hand and wrist movements. Each of the finalised video recordings was two to six minutes long.

2.3. Procedure

In the beginning of the first meeting, a 25-min introductory lecture was given. The lecture included general information regarding procedures for performing risk assessments. Special emphasis was put on the quantification of work task exposure in the dimensions of intensity, frequency and duration of work task. The lecture also addressed the increased demands made by the Swedish Work Environment Authority (SWEA) concerning ergonomic risk assessments (SWEA, 2012), where a paragraph (§4) in the present Ergonomic provisions has been added in comparison to

Table 1
Characteristics of the observers (n = 21).

Observer characteristics		
Age, mean (range)	51 (40–64)	
Women, n (%)	20 (95)	
Years of work experience within physical ergonomics, mean (range)	14 (4–26)	
Client Company Sectors, n (%)		
	Industry	16 (76)
	Office	15 (71)
	Service and Trade	4 (19)
	Healthcare	7 (33)
	Other ^a	3 (14)
Frequency of risk assessment assignments, n (%)		
	Once a week	4 (19)
	Once a month	8 (38)
	Once every three months	5 (23)
	Once every six months	1 (5)
	Once a year	2 (10)
	Less than once a year	1 (5)

^a Other sectors; e.g. the Swedish Armed Forces and different municipal sectors.

Table 2
Description of work tasks used in the study.

Task activity	Hours per workday	Weight of handled goods (kg)	Environment, physical factors	Ratings of discomfort (BORG-scale)	Work demands and control ^a
Unpack groceries to shelves in a supermarket store	just over 4	2	Good	3	Partly autonomy
Put nets around roasts at a slaughterhouse	just over 4	2.5–4.5	Cold, wet, noisy	4	Autonomy group
Throw small boxes into containers (postal sorting work)	just over 2	3	Cold during winter, warm during summer, noisy, difficulty concentrating	3–4	Controlled
Put bundles of letters into boxes (postal sorting work)	6	2	Cold during winter, warm during summer, noisy, difficulty concentrating	3–4	Controlled
Deboning meat at a slaughterhouse	7	3–4	Cold, wet, noisy, sharp knives	3–4	Autonomy group
Engine assembly	just under 3	2	Good	2.5	Controlled
Hair cutting	just over 4	1	Good	3	Autonomy
Lavatory cleaning	5	1	Good	2	Partly autonomy
Supermarket cashier work	7	1–5	Good	3	Controlled
Cleaning stairs	just under 4	1	Usually good, sometimes cold	3	Partly autonomy

^a Partly autonomy: The worker controls the work task, but is limited in time and by obligations of other work tasks included in the work. Autonomy group: a group of employees control and divide work tasks within the group. Controlled: The work task is completely controlled in time by work instructions and in space by the physical design of the workplace. Autonomy: The worker controls the work himself/herself as if self-employed.

the previous version, emphasizing the employers' responsibility to perform risk assessments of the work environment with regards to work postures and movements, manual handling and repetitive work, taking into account the duration, frequency and amplitude of the different exposures.

During the risk assessments, the observers watched the video recordings on individual laptop computers and were allowed to pause or repeat the playback as needed. Documents with complementary information about the different work tasks, such as duration of work task during the work day, pause and rest schedules, weight of handled goods, other physical factors, and ratings of discomfort on Borg's CR10-scale (Borg, 1998), as well as work demands and control were provided for each work task (Table 2). The maximum assessment time for each work task was set at 20 min.

Using a separate protocol for each work task, the observers rated the risk for MSD in eight specific body regions: neck, lower back, right and left shoulders, arms/elbows, and wrists/hands. Finally they rated the global risk of the whole work task.

Risk was rated using a three-stage scale: red (high risk), yellow (moderate risk), and green (low risk). Instructions were given using the wording in the Ergonomics Provisions from the SWEA, where a red rating is to be given when the exposure in the work task is judged as unacceptable, all or almost all workers will likely develop MSDs, and immediate action needs to be taken. A yellow rating bare the meaning that several workers may develop MSDs, and the work task needs to be investigated further, and a green rating corresponds to an acceptable exposure (SWEA, 2012) where none or sporadic individuals may develop MSDs. These instructions were chosen taking into consideration that all the ergonomists were familiar with the SWEA provisions.

In contrast to when systematic observational methods are used, the risk levels that the ergonomists rated were not based on a calculated score from ratings of different exposure parameters. The observers made their observational risk assessment solely based on his/her experience and knowledge regarding work-related exposures and risk for MSD, without using an explicit method.

Intra-observer reliability was assessed by having nine of the ergonomists carry out the whole procedure a second time within a minimal three week interval.

2.4. Statistical analysis

Calculation of inter-observer reliability was based on assessments of the 21 observers' initial assessments. Calculation of intra-

observer reliability was based on the first and second ratings of the nine observers who repeated their assessments. Statistics for each of the 10 work tasks regarding the eight body regions, as well as for the total risk assessment, was calculated.

Several statistics for reliability were calculated to enable comparisons with other studies. Agreement (%) was calculated as the number of rating pairs in agreement divided by the total number of rating pairs. To take the agreement due to chance into account, percent agreement will always be presented together with other parameters, for example, kappa statistics (Cohen, 1960). For the intra-observer reliability, Cohen's kappa was calculated for each of the nine observers, and then the mean value of these kappa values was used (Cohen, 1960). Since Cohen's kappa is only applicable when two raters are used or when test–retest reliability is evaluated, a kappa that was pairwise averaged over all pairs was calculated for the inter-observer reliability. This was done in the way suggested by Davies and Fleiss (1982), where the expected agreement, P_e , in Cohen's kappa formula for each pairwise comparison, $k = (Po - P_e) / (1 - P_e)$, is substituted with the average P_e of all pairs. Since the risk ratings represent ordinal data (low, moderate and high risk) and Cohen's unweighted kappa does not distinguish minor from major discrepancies in ratings, the linearly weighted kappa (Cohen, 1968; Warrens, 2012) was also computed and averaged in the same way as the unweighted kappa (Davies and Fleiss, 1982; Hallgren, 2012; Sawa and Morikawa, 2007). The intraclass correlation (ICC), two-way absolute agreement method 2.1, according to Shrout and Fleiss (1979), was also computed to facilitate comparisons with other studies (Comper et al., 2012; David et al., 2008; Paulsen et al., 2015; Spielholz et al., 2008; Stephens et al., 2006). ICC is mostly applicable for continuous data but has been used in these reliability studies on ordinal data. Also Kendall's coefficient of concordance (KCC) was computed. KCC is a non-parametric relative to ICC that is applicable with ordinal data (McDowell, 2006).

Landis and Koch's (1977) table was used for interpretation of kappa. A kappa value higher than or equal to 0.41, which indicates at least moderate agreement, was considered as indicating acceptable reliability.

The statistical computations were carried out using scripts written in MATLAB version 8.5 (MathWorks Inc., Natick, MA, USA), the output parameters of which, for small samples, were compared and found to agree with corresponding parameters of the statistical software R or SPSS. MATLAB was used in order to obtain time effective analyses; since there were no functions for multi-observer linearly weighted kappa in R or SPSS.

3. Results

3.1. Inter-observer reliability

Altogether, 1680 body part risk assessments (21 ergonomists, eight body regions, 10 work tasks) were made (Fig. 1), whereof 404 assessments (21%) were rated as high risk (red), 857 (45%) moderate risk (yellow) and 629 (33%) low risk (green). Of the 210 global risk assessments, 76 (36%) were rated as high risk, 98 (47%) as moderate risk, and 36 (17%) as low risk.

Both high and low risk ratings of the global risk were present in seven out of ten work tasks. The body regions where both high and low risk ratings were least frequent were the lower back, followed by the left shoulder. For the lower back, five out of 10 work tasks were classified as both high and low risk, while the same was true for six out of 10 work tasks for left shoulder. For the other body regions, at least seven of ten work tasks were classified with both high and low risk ratings (Fig. 1). All ergonomists rated “engine assembly” as low or moderate risk for all body regions. That was the only work task with no high risk rating for any of the eight body regions.

The average weighted kappa for the inter-observer reliability was between 0.12 and 0.32 for the different body regions and 0.31 for the global risk. The body regions with the highest reliability scores were the left shoulder, neck and low back (Table 3).

3.2. Intra-observer reliability

The intra-observer reliability parameters representing assessments from the nine ergonomists who performed a second assessment are presented in Table 4. The agreement of the global risk assessments was 61% (weighted kappa 0.41).

4. Discussion

In this study, the inter-observer reliability of risk estimates for the eight body regions showed weighted kappas of 0.12–0.32. Hence, the reliability risk estimates of all body regions were in the range of slight-to-fair, and all fell below the ‘moderate reliability’ criteria of 0.41 that Landis and Koch (1977) suggested. The inter-rater reliability regarding the global risk (weighted kappa = 0.32) indicated a fair reliability according to the same criteria (weighted kappa 0.21–0.41). As expected, the intra-observer reliability was somewhat higher 0.62 for low back, 0.23–0.4 for the other body regions, and 0.41 for the global risk.

In the present study, the ergonomists’ risk assessment was based on their own knowledge and experience. One would assume that when ergonomists use specific, more elaborate methods, it will probably lead to a higher level of reliability. Takala et al. (2010) wrote in their evaluation of 30 observational methods that information regarding the methods’ reliability was limited. However, there are some reported reliability studies in the literature. Table 5 shows the results of such studies, but the various studies used different statistics methods, making a comparison hard to perform.

The reliability found in the present study is generally lower than in the tabulated observational methods. In the present study, the aim was to mimic the way in which risk assessments often are performed by ergonomists in the OHS, which – as stated in the introduction – often is done solely by means of observation, based on the ergonomist’s knowledge and experience, without the use of an explicit method. This was done using a three level green-yellow-red scale, where the scale steps are equivalent to low – medium – high risk of MSD, as they are described in the Swedish Ergonomic Provisions from SWEA (2012). This “traffic-light” model

is well known to Swedish ergonomists (Eliasson et al., unpublished manuscript). This model is also used with similar definitions, for risk indexes in several widely spread observational methods, such as; Assessment of Repetitive Tasks of the upper limbs (the ART tool) (Ferreira et al., 2009), Hand Activity Level (HAL) (Armstrong, 2006), and the hand-arm risk assessment method (HARM) (Douwes and de Kraker, 2009, 2014). Differences between the observers’ perception of the body postures and movements, and in their transformation to risk estimates contribute, of course, to the differences in assessed risks. These differences in perception of risk can be reduced by using systematic risk assessment methods, where the risk level is defined based on a calculated score from ratings of different exposure parameters.

There are several additional possibly contributing explanations for the low levels of reliability. One other possible reason for both the low inter- and intra-observer reliability could be that the observers concentrated and based their assessment on different aspects and time periods within the two-to-six minute long video recordings. Additional explanations may be found in the differences in experience in risk assessment among the observers and also in their experience of different work sectors. A further possible contributing factor could be the difficulty in assessing postures and movements from video recordings in comparison with live observations (Leskinen et al., 1997). However, Mathiassen et al. (2013) found that posture assessments conducted using observation should be based on video recordings because this permits repeated ratings of the same work sequence. In the present study, several synchronised views were used to create a realistic situation. This enabled the observers to analyse the work postures and movements from different angles, which gave a more comprehensive picture of the work situation.

General difficulties in the assessment of postures and movements can contribute to divergent risk assessments. Earlier studies indicated that small body movements, such as hand and wrist activities, are harder to assess using observation than are larger body part movements, such as movements of the back (Takala et al., 2010), as seen in this study.

Many different parameters of reliability have been used in studies of observational methods. We included the most commonly used ones, in order to be able to compare with previous studies (Table 5). In addition to the linearly weighted kappa, which is considered the most suitable one (Warrens, 2012), Cohen’s unweighted kappa (Cohen, 1968) can be used, but is most suitable for nominal data and it does not differ between one- and two-category differences, while with the linearly weighted kappa, a two-category difference is given double the weight of a one-category difference. ICC (Shrout and Fleiss, 1979) is very similar to a quadratically weighted kappa (Fleiss and Cohen, 1973; Hallgren, 2012), and it often recommended for use in multi-observer comparisons since it is included in all statistical packages. However, we consider both ICC and quadratically weighted kappas to weight one-category differences too low. In this study, a difference between, for example, moderate and high risk was weighted to 0.25 with ICC and 0.5 with linearly weighted kappa, which we consider more reasonable. Another advantage of a linearly weighted kappa is that scales with different steps can be “fairly” compared, which can hardly be done with quadratically weighted or unweighted kappas (Brenner and Klietsch, 1996; Warrens, 2012).

A strength of the present study is that there were as many as 21 observers (for the inter-reliability part), and all were experienced ergonomists. It is common in reliability studies that fewer observers than this participate (Comper et al., 2012; David et al., 2008; Paulsen et al., 2015; Spielholz et al., 2008).

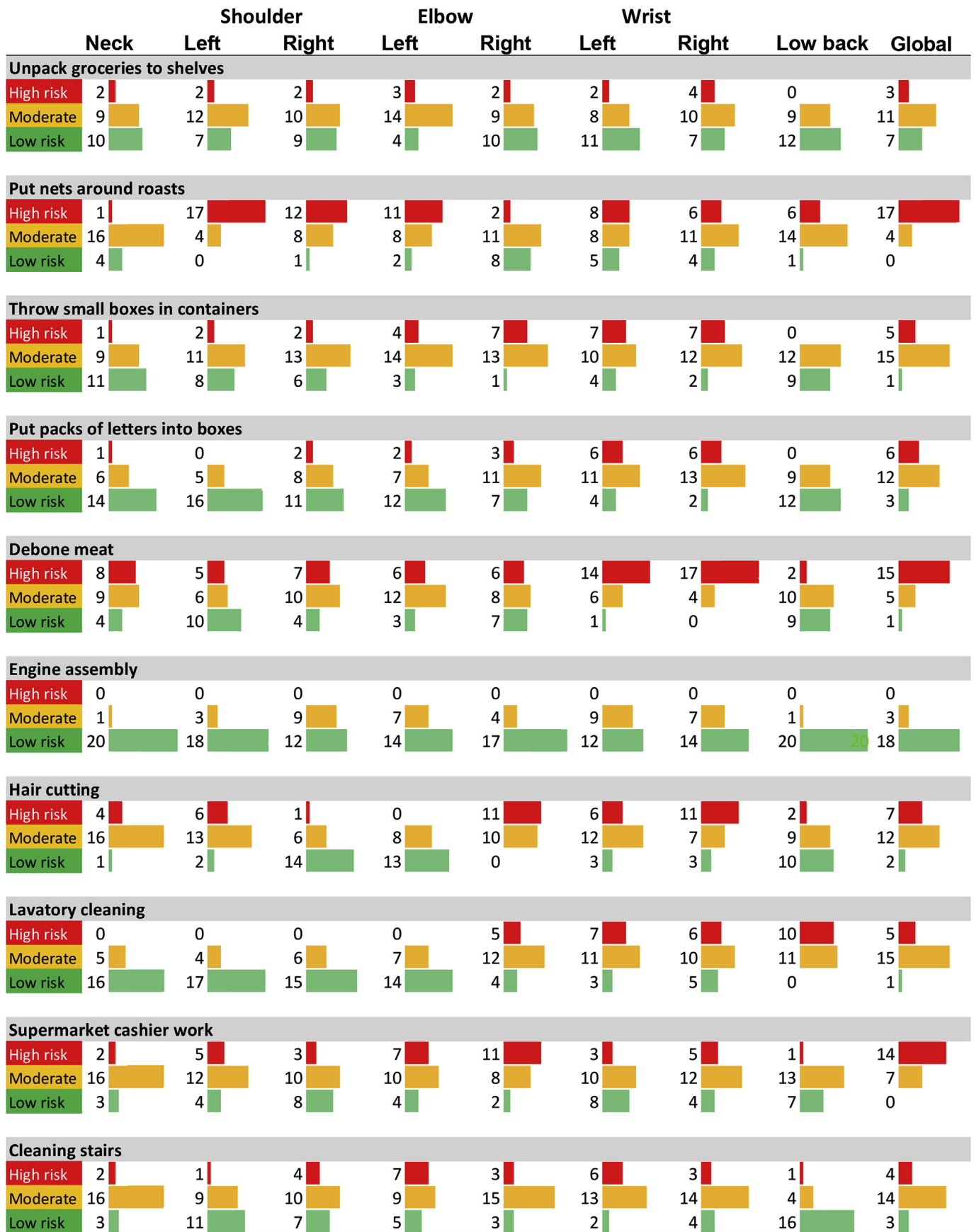


Fig. 1. Risk assessment of 10 different work tasks. The bars represent the number of observers' risk assessments for specific body regions, and, in the rightmost column, the global risk. Only the observers' initial assessments are included.

Table 3

Inter-observer reliability of the risk assessments in terms of agreement (%), Cohens kappa (κ), weighted kappa averaged over pairs (κ_w), intraclass correlation (ICC) and Kendall's coefficient of concordance (KCC) (n = 21).

	Inter-observer reliability*				
	%	κ	κ_w	ICC	KCC
Neck	56	0.24	0.27	0.33	0.43
Right shoulder	44	0.13	0.18	0.27	0.36
Left shoulder	51	0.23	0.32	0.43	0.50
Right elbow	40	0.07	0.12	0.19	0.30
Left elbow	43	0.09	0.15	0.24	0.38
Right wrist	44	0.11	0.16	0.24	0.32
Left wrist	44	0.12	0.18	0.28	0.41
Low back	52	0.19	0.26	0.37	0.40
Global risk	53	0.25	0.31	0.43	0.50

*The ergonomists' initial assessments.

Table 4

Intra-observer reliability of the risk assessments in terms of agreement (%), mean Cohens kappa (κ), weighted kappa (κ_w), intraclass correlation (ICC) and Kendall's coefficient of concordance (KCC) (n = 9).

	Intra-observer reliability				
	%	κ	κ_w	ICC	KCC
Neck	59	0.28	0.35	0.45	0.77
Right shoulder	53	0.21	0.30	0.43	0.73
Left shoulder	57	0.29	0.38	0.51	0.75
Right elbow	50	0.17	0.23	0.31	0.66
Left elbow	64	0.36	0.40	0.48	0.73
Right wrist	51	0.13	0.20	0.29	0.67
Left wrist	54	0.23	0.28	0.36	0.71
Low back	76	0.58	0.62	0.69	0.86
Global risk	61	0.32	0.41	0.54	0.77

4.1. Limitations of the study

In real situations, the ergonomists interview employees while on site at the workplace. This limitation in the present study was addressed by supplying the observers with written information on the different work tasks. In general, reliability is dependent on the heterogeneity of the study sample (de Vet et al., 2006). The included work tasks were selected due to their heterogeneity and represent different levels of risk for MSD, as well as differences in handled weights, cycle times, and level of standardisation. For practical reasons, there were not more than 10 different tasks included. The ergonomists agreed to different degrees in the various work tasks, so the results may have been somewhat different if other or more tasks were included. The engine assembly task was the one of most "green" ratings, and it had the most agreement. If more work tasks were "very green" or "very red", the global risk reliability would have been higher.

Table 5

Overview of inter-observer reliability studies of methods used mainly for the assessment of upper extremity risks and comparison with the present study. Statistics presented with Cohen kappa (κ), Kendall's coefficient of concordance (KCC), intraclass correlation (ICC) and weighted kappa (κ_w).

Observational method	Inter-observer reliability	Present study	Reference
Quick Exposure Check (QEC) Phase 1	κ 0.17–0.42	κ 0.09–0.25	David et al. (2008)
Quick Exposure Check (QEC) Phase 2	KCC 0.6–0.76	KCC 0.32–0.5	David et al. (2008)
Quick Exposure Check (QEC) Brazilian–Portuguese version	ICC 0.62–0.86	ICC 0.29–0.69	Comper et al. (2012)
Hand Activity Level (HAL)	κ_w 0.34 ^a	κ_w 0.31 (global) ICC 0.43	Spielholz et al. (2008)
Strain Index score (SI)	κ_w 0.41 ^a	κ_w 0.31 (global) ICC 0.43	Spielholz et al. (2008)
Strain Index (SI)	ICC 0.56	ICC 0.43	Stephens et al. (2006)
OCRA checklist score	ICC 0.59	ICC 0.43	Paulsen et al. (2015)
OCRA checklist risk	ICC 0.54	ICC 0.43	Paulsen et al. (2015)

^a Probably the quadratically weighed kappa was used. Quadratically weighted kappa is more comparable with ICC compared to with linearly weighted kappas.

In theory, the assessments of a group of observers may be performed with perfect reliability but with low validity, i.e. if all observers do the same but incorrect assessments. However when there is a low reliability (low agreement), there cannot be a high validity, since a high reliability is a necessary but not sufficient condition for high validity (American Educational Research Association, 1985). In the present study we could not analyse validity. Since it was not possible to form any "correct" risk estimates (gold standards). However, according to the theory described above, and with the low reliability, if there had been a way to obtain a gold standard, the concurrent validity would also be low. To investigate the predictive validity longitudinal data and large cohorts are required (Mokkink et al., 2010), but again with the low reliability in this study, also the predictive validity would be low.

There are some examples of observational methods that have been evaluated regarding predictive validity. These methods are mainly focused on exposures towards the hand and lower arms (Takala et al., 2010). Both index from ACGIH TLV for hand activity and Strain index has been associated with MSD's in the distal part of the upper extremities. (Bonfiglioli et al., 2013; Kapellusch et al., 2014; Spielholz et al., 2008).

4.2. Practical implications

As the results of this study show, to just observe and assess risk assessment of repetitive work without any systematic methodology is likely to give unreliable results. Even when comparisons are made with previous assessments made by the same ergonomist the reliability are likely to be rather low. Although there are also reports of limited reliability for systematic observational methods (Takala et al., 2010) and different methods may show different levels of risk (Chiasson et al., 2012), generally, systematic methods show a higher level of reliability. Since earlier studies have discussed (Takala et al., 2010; Eliasson et al., unpublished manuscript) that there is a lack of use of systematic observational methods among ergonomists, there is a continuing need for education in this important topic, and also a need for work regarding the dissemination of knowledge concerning risk assessment. Reports may be consulted with for decisions regarding risk assessment methods suitable for the particular exposure type (Neumann, 2007; Takala et al., 2010; Palm et al., 2014). Technical measurements can also be used in risk assessments. Development of cheap and feasible techniques, such as accelerometers, enable practitioners to perform technical measurements of movements and postures (Dahlqvist et al., 2016) in combination with observational methods. In terms of precision, inclinometers were more cost-efficient regarding assessment of trunk and upper arm inclination compared with observation (Trask et al., 2014). Today, there are recommendations for wrist angular velocities (Nordander et al., 2013), in addition to

ongoing longitudinal research exploring dose–response relationships for arm and back postures. There is, however, still a lack of reference values and threshold limits to guide the practitioner in the interpretation of data in a risk assessment. Technical measurements have a high reliability and may be used to compare the exposure between workplaces or assess the effect on the exposure before and after an intervention.

Ergonomists often make risk assessments by themselves. To increase the reliability of the risk estimate, one approach could be to use several observers and to use the mean value of their assessments, or to let the project leading ergonomist be consulted for assessment of her/his colleagues (as it may be hard to average three category ratings). Mathiassen et al. (2012) found that in the assessment of arm postures, it is more cost effective to use a higher number of observers as compared to having one observer view more videos. Within an OHS organisation, this could be achieved by having several ergonomists observe video recordings of the work task to be assessed.

In future research, it might be of interest to explore the possibility to include other types of information in the risk assessment. Such additional information could include statistics on MSD incidents the workplace and at other workplaces with similar levels of exposure. However, this kind of information can be precarious to evaluate due to healthy worker effects, meaning that in a workplace with harmful exposures, workers with symptoms typically have already quit due to health reasons (Shah, 2009). This needs to be investigated further.

5. Conclusion

Assessments of risks for MSDs using observation without an explicit method have low, non-acceptable levels of reliability. It is therefore recommended to use systematic risk assessment methods to a higher degree.

Acknowledgement

The Swedish Research Council for Health, Working Life and Welfare (FORTE project no. 1212–1202) supported this study. In addition to the authors of this article, the project team consisted of Ida-Märta Rhen, Katarina Kjellberg, Natalja Balliu, Liv Egnell and Per Lindberg, who all participated in fruitful discussions and support (Liv and Natalja) with statistical computations.

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1985. Standards for Educational and Psychological Testing. American Psychological Association, Washington, DC.
- Armstrong, T., 2006. The ACGIH TLV for hand activity level. In: Marras, W.S., Karwowski, W. (Eds.), *Fundamentals and Assessment Tools for Occupational Ergonomics*, vol. 41. CRC Press, Boca Raton, Florida, pp. 1–14.
- Bonfiglioli, R., Mattioli, S., Armstrong, T.J., Graziosi, F., Farioli, A., Violante, F.S., 2013. Validation of the ACGIH TLV for hand activity level in the OCTOPUS cohort: a two-year longitudinal study of carpal tunnel syndrome. *Scand. J. Work Environ. Health* 39, 155–163.
- Bongers, P.M., Ijmker, S., van den Heuvel, S., Blatter, B.M., 2006. Epidemiology of work related neck and upper limb problems: psychosocial and personal risk factors (Part I) and effective interventions from a bio behavioural perspective (Part II). *J. Occup. Rehabil.* 16, 272–295.
- Bovenzi, M., 2006. Health risks from occupational exposures to mechanical vibration. *La Med. del Lav.* 97, 535–541.
- Borg, G., 1998. Borg's Perceived Exertion and Pain Scales. Human Kinetics, Champaign, IL.
- Brenner, H., Kliebsch, U., 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiol. Camb. Mass.* 7, 199–202.
- Chiasson, M.-È., Imbeau, D., Aubry, K., Delisle, A., 2012. Comparing the results of eight methods used to evaluate risk factors associated with musculoskeletal disorders. *Int. J. Ind. Ergon.* 42, 478–488.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46.
- Cohen, J., 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* 70 (4), 213–220.
- Comper, M.L., Costa, L.O., Padula, R.S., 2012. Clinimetric properties of the Brazilian–Portuguese version of the quick exposure check (QEC). *Rev. Bras. Fisioter. (Sao Carlos (Sao Paulo, Brazil))* 16, 487–494.
- Dahlqvist, C., Hansson, G.-H., Forsman, M., 2016. Validity of a small low-cost triaxial accelerometer with integrated logger for uncomplicated measurements of postures and movements of head, upper back and upper arms. *Appl. Ergon.* 55 (July), 108–116.
- David, G., Woods, V., Li, G., Buckle, P., 2008. The development of the Quick Exposure Check (QEC) for assessing exposure to risk factors for work-related musculoskeletal disorders. *Appl. Ergon.* 39, 57–69.
- Davies, M., Fleiss, J.L., 1982. Measuring agreement for multinomial data. *Biometrics* 38, 1047–1051.
- de Vet, H.C., Terwee, C.B., Knol, D.L., Bouter, L.M., 2006. When to use agreement versus reliability measures. *J. Clin. Epidemiol.* 59, 1033–1039.
- Dempsey, P.G., McGorry, R.W., Maynard, W.S., 2005. A survey of tools and methods used by certified professional ergonomists. *Appl. Ergon.* 36, 489–503.
- Douwes, M., de Kraker, H., 2009. Hand Arm Risk assessment Method (HARM), a new practical tool. In: 17th World Congress on Ergonomics. International Ergonomics Association.
- Douwes, M., de Kraker, H., 2014. Development of a non-expert risk assessment method for hand-arm related tasks (HARM). *Int. J. Ind. Ergon.* 44, 316–327.
- Eliasson, K., Lind, C., Nyman, T., 2016. Ergonomics Risk Assessment: Tool Use and Processes. Unpublished manuscript.
- European Council, 1989. Council Directive 89/391/EEC of 12 June 1989 on the Introduction of Measures to Encourage Improvements in the Safety and Health of Workers at Work.
- Ferreira, J., Gray, M., Hunter, L., Birtles, M., Riley, D., 2009. Development of an Assessment Tool for Repetitive Tasks of the Upper Limbs (ART). Derbyshire.
- Fleiss, J.L., Cohen, J., 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* 33 (3), 613–619.
- Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials Quant. Methods Psychol.* 8, 23–34.
- Hägg, G.M., 2003. Corporate initiatives in ergonomics—an introduction. *Appl. Ergon.* 34, 3–15.
- Kapellusch, J.M., Gerr, F.E., Malloy, E.J., Garg, A., Harris-Adamson, C., Bao, S.S., et al., 2014. Exposure–response relationships for the ACGIH threshold limit value for hand-activity level: results from a pooled data study of carpal tunnel syndrome. *Scand. J. Work Environ. Health* 44, 610–620.
- Ketola, R., Toivonen, R., Viikari-Juntura, E., 2001. Interobserver repeatability and validity of an observation method to assess physical loads imposed on the upper extremities. *Ergonomics* 44, 119–131.
- Koningsveld, E.A.P., Dul, J., Van Rhijn, G.W., Vink, P., 2005. Enhancing the impact of ergonomics interventions. *Ergonomics* 48, 559–580.
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- Lang, J., Ochsmann, E., Kraus, T., Lang, J.W.B., 2012. Psychosocial work stressors as antecedents of musculoskeletal problems: a systematic review and meta-analysis of stability-adjusted longitudinal studies. *Soc. Sci. Med.* 75, 1163–1174.
- Leskinen, T., Hall, C., Rauas, S., Ulin, S., Tonnes, M., Viikari-Juntura, E., Takala, E.P., 1997. Validation of portable ergonomic observation (PEO) method using optoelectronic and video recordings. *Appl. Ergon.* 28, 75–83.
- Lind, C., Rose, L., 2016. Shifting to proactive risk management: risk communication using the RAMP tool. *Agron. Res.* 14, 513–524.
- Lowe, B.D., 2004. Accuracy and validity of observational estimates of wrist and forearm posture. *Ergonomics* 47, 527–554.
- Mathiassen, S.E., Liv, P., Wahlström, J., 2012. Cost-efficient observation of working postures from video recordings – more videos, more observers or more views per observer? *Work* 41, 2302–2306.
- Mathiassen, S.E., Liv, P., Wahlström, J., 2013. Cost-efficient measurement strategies for posture observations based on video recordings. *Appl. Ergon.* 44, 609–617.
- McDowell, I., 2006. *Measuring health: a guide to rating scales and questionnaires*. Oxford University Press.
- Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C., 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* 63, 737–745.
- Neumann, W.P., 2007. *Inventory of Human Factors Tools and Methods – a Work System Design Perspective*. Human Factors Engineering Lab. Technical Report.
- Nordander, C., Ohlsson, K., Akesson, I., Arvidsson, I., Balogh, I., Hansson, G.A., Stromberg, U., Rittner, R., Skerfving, S., 2013. Exposure–response relationships in work-related musculoskeletal disorders in elbows and hands – a synthesis of group-level data on exposure and response obtained using uniform methods of data collection. *Appl. Ergon.* 44, 241–253.
- Palm, P., Eliasson, K., Lindberg, P., Hägg, G., 2014. Belastningsergonomisk riskbedömning –Vägledning och metoder. Rapport nr 1/2014. Arbets- och miljömedicin, Uppsala.
- Palmer, K.T., Harris, E.C., Coggon, D., 2007. Carpal tunnel syndrome and its relation to occupation: a systematic literature review. *Occup. Med. (Lond)* 57, 57–66.
- Palmer, K.T., Smedley, J., 2007. Work relatedness of chronic neck pain with physical

- findings – a systematic review. *Scand. J. Work Environ. Health* 33, 165–191.
- Paulsen, R., Gallu, T., Gilkey, D., Iireiser, R., Murgia, L., Rosecrance, J., 2015. The inter-rater reliability of Strain Index and OCRA Checklist task assessments in cheese processing. *Appl. Ergon.* 51, 199–204.
- Punnett, L., Wegman, D.H., 2004. Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. *J. Electromyogr. Kinesiol.* 14, 13–23.
- Putz-Anderson, V., et al., 1997. In: Bernard, B.P. (Ed.), *Musculoskeletal Disorders and Workplace Factors – a Critical Review of Epidemiologic Evidence for Work-related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back*. U.S. Department of health and human services, Public health service centers for disease control and prevention, National Institute for Occupational Safety and Health. Publication no.: 97–141.
- Sawa, J., Morikawa, T., 2007. Interrater reliability for multiple raters in clinical trials of ordinal scale. *Drug Inf. J.* 41, 595–605.
- Shah, D., 2009. Healthy worker effect phenomenon. *Indian J. Occup. Environ. Med.* 13, 77–79.
- Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 86, 420.
- Spielholz, P., Bao, S., Howard, N., Silverstein, B., Fan, J., Smith, C., Salazar, C., 2008. Reliability and validity assessment of the hand activity level threshold limit value and strain index using expert ratings of mono-task jobs. *J. Occup. Environ. Hyg.* 5, 250–257.
- Spielholz, P., Silverstein, B., Morgan, M., Checkoway, H., Kaufman, J., 2001. Comparison of self-report, video observation and direct measurement methods for upper extremity musculoskeletal disorder physical risk factors. *Ergonomics* 44, 588–613.
- Stephens, J.P., Vos, G.A., Stevens, E.M., Moore, J.S., 2006. Test-Retest repeatability of the strain index. *Appl. Ergon.* 37 (3), 275–281.
- Stevens, E., Gordon, A., Stephens, J.-P., Moore, S., 2004. Inter-rater reliability of the Strain index. *J. Occup. Environ. Hyg.* 1 (11), 745–751.
- Swedish Work Environment Authority, 2012. *Belastningsergonomi: Arbetsmiljöverkets föreskrifter och allmänna råd om belastningsergonomi (Physical Ergonomics: Provisions of the Swedish Work Environment Authority together with general recommendations on the implementation of the Provisions)*. AFS 2012:2. [in Swedish]. Arbetsmiljöverket, Stockholm.
- Takala, E.P., Pehkonen, I., Forsman, M., Hansson, G.A., Mathiassen, S.E., Neumann, W.P., Sjogaard, G., Veiersted, K.B., Westgaard, R.H., Winkel, J., 2010. Systematic evaluation of observational methods assessing biomechanical exposures at work. *Scand. J. Work Environ. Health* 36, 3–24.
- Trask, C., Mathiassen, S.E., Wahlström, J., Forsman, M., 2014. Cost-efficient assessment of biomechanical exposure in occupational groups, exemplified by posture observation and inclinometry. *Scand. J. Work Environ. Health* 40, 252–265.
- van Rijn, R.M., Huisstede, B.M., Koes, B.W., Burdorf, A., 2009a. Associations between work-related factors and specific disorders at the elbow: a systematic literature review. *Rheumatol. Oxf. Engl.* 48, 528–536.
- van Rijn, R.M., Huisstede, B.M., Koes, B.W., Burdorf, A., 2009b. Associations between work-related factors and the carpal tunnel syndrome – a systematic review. *Scand. J. Work Environ. Health* 35, 19–36.
- van Rijn, R.M., Huisstede, B.M., Koes, B.W., Burdorf, A., 2010. Associations between work-related factors and specific disorders of the shoulder – a systematic review of the literature. *Scand. J. Work Environ. Health* 36, 189–201.
- Warrens, M.J., 2012. Conditional inequalities between Cohen's kappa and weighted kappas. *Stat. Methodol.* 10, 14–22.
- Wells, R.P., Neumann, W.P., Nagdee, T., Theberge, N., 2013. Solution building versus problem convincing: ergonomists report on conducting workplace assessments. *IIE Trans. Occup. Ergon. Hum. Factors* 1, 50–65.
- Whysall, Z.J., Haslam, R.A., Haslam, C., 2004. Processes, barriers, and outcomes described by ergonomics consultants in preventing work-related musculoskeletal disorders. *Appl. Ergon.* 35, 343–351.