Chapter 7

Inter-rater reliability of a video-analysis method measuring low-back load in a field situation.

P. Coenen
I. Kingma
C.R. Boot
P.M. Bongers
J.H. van Dieën

Applied Ergonomics, 2013,
44(5): 828-834

ABSTRACT

Valid and reliable low-back load assessment tools that can be used in field situations are needed for epidemiologic studies and for ergonomic practice. The aim of this study was to assess the inter-rater reliability of a low-back load video analysis method in a field setting.

Five raters analyzed 50 work site manual material handling tasks of 14 workers. Peak and mean moments at the level of L5S1, and segment angles were obtained using the video analysis method. Intra-class correlation coefficients (ICCs) and median standard deviations across raters were calculated.

ICCs revealed excellent inter-rater reliability (>0.9) for peak and mean moments, ICCs of segment angles were variable. Median standard deviations showed relatively small inter-rater variance for moments (standard deviation <10 Nm) and segment angle variation ranging from 0° to 20°. The proposed video analysis method provides a reliable tool for obtaining low-back loads from occupational field tasks.

INTRODUCTION

High low-back loads that may occur at work (e.g. during lifting, pushing and pulling of objects or working in awkward body positions) are associated with low-back pain (LBP; e.g., Marras et al., 2010; van Dieën et al., 1999). These associations have often been confirmed in epidemiological studies using self-reported exposures or field observations (da Costa & Vieira, 2010; Griffith et al., 2012; Lötters et al., 2003). However, other epidemiological studies did not find support for the association between high low-back loads and LBP, possibly as a result of the lack of appropriate measurement designs (Bakker et al., 2009). Therefore, valid and reliable low-back load assessment methods that can be applied in field settings are needed. Three types of measurement methods can be adopted: self-reports, observational techniques and direct measurement techniques (Burdorf, 2010; David, 2005). Although self-reports are highly efficient, they are assumed to be less reliable than observational techniques and direct measurements (Balogh et al., 2004; Hansson et al., 2001). On the other hand, direct measurement techniques (e.g., measuring muscle activity or body posture recordings using marker tracking or goniometry) are much more accurate but difficult to apply in large scale field studies. In field measurements of low-back load, there thus seems to be a trade-off between efficiency (in terms of time, money and resources) and accuracy. Besides, it can be argued that crude observational low-back exposure measures (e.g., the number of lifts, time spent in a flexed trunk position) provide less detailed information on low-back load than dose metrics (i.e., low-back moments), since different exposures (e.g., lifting and bending) affect the same dose. Therefore, dose-estimates can provide more insight into the etiology of LBP (Wells et al., 2004) and these metrics are more predictive of future LBP than postural exposure measures (Coenen et al., 2013b).

Video-based methods using postural exposure data in biomechanical models to calculate low-back load dose estimates have been shown to be a promising category of observational techniques (e.g., Chang et al., 2010; Coenen et al., 2011; Norman et al., 1998; Potvin, 1997; Sutherland et al., 2008) in the assessment of low-back load metrics such as static (Neumann et al., 2001b), cumulative (Sutherland et al., 2008) or peak low-back moments (Norman et al., 1998). Furthermore, these coding systems allow raters with minimal training and minor use of equipment to collect occupational low-back load data. High inter-rater agreement has been found when using these kinds of models to calculate cumulative low-back moments (Cann et al., 2008; Sullivan et al., 2002). However, testing of these models was only performed in laboratory situations or in mock-ups of field situations, whereas, applicability of these methods for epidemiological studies or in ergonomic practice can best be assessed when applied to actual field situations. The aim of the present study therefore was to test the inter-rater reliability of a low-back load video analysis method in a field setting. The model that will be tested in our study has been validated against a lab-based reference method (Coenen et al., 2011) and inter-rater

reliability has been assessed in a laboratory situation (Xu et al., 2011). Although these authors suggest that the method might be valid and reliable in field studies, reliability has not yet been assessed in field settings.

MATERIAL AND METHODS

Data collection

Videos of a wide range of manual materials handling (MMH) tasks were selected from the SMASH cohort that has been described before (Ariëns et al., 2001; Hoogendoorn et al., 2000a). Briefly, in this cohort, risk factors of musculoskeletal disorders were studied in workers from various industrial and service branches, for example, in the metal, chemical, pharmaceutical, food and wood construction industry; waste processing, insurance and distribution companies. The SMASH study consists of a baseline measurement, assessing physical load at the workplace, and baseline and three year follow-up assessment of musculoskeletal symptoms. For the assessment of physical work load, 5–15 min of video recordings at the workplace were taken at four moments during the course of one day. During these recordings, researchers handling the camera were instructed to take a sagittal plane view as much as possible. For all MMH tasks during these 15 min, external forces at the hands were measured using force transducers (during pushing and pulling) or weighing scales measuring mass of the external load (during lifting). Afterward, videos were systematically observed during which MMH tasks, i.e. lifting, pushing and pulling tasks during which external forces are exerted on the hands, were identified. Fifty video fragments were selected representing tasks (38 lifting, 6 pushing and 6 pulling tasks), executed by 14 workers of 10 particular companies. Rather than randomly selecting, we carefully selected these tasks, in order to obtain a wide range of tasks, work postures, task asymmetry, physical workloads and image quality and camera angle relative to the sagittal plane of the subject. Thus, we also included tasks that had not been recorded optimally, e.g. due to occlusion of the view by another worker or with a large angle between the camera plane and the sagittal plane of the subject. The selected workers were 31.9(8.3) years of age and seven workers were female. Six workers reported LBP at baseline. External forces at the hands measured during these tasks were on average 66 (80) N and ranged from almost 0 N to 368 N.

Five raters were recruited among students of the Amsterdam School of Health Professions. Three of them were third year physical therapy students and two of them were fourth year occupational therapy students. The raters were 22.2 (1.8) years of age and had substantial knowledge on kinesiology. After participating in an extensive learning and practice session in which the raters were briefed regarding the purpose of the study and were familiarized with the software, raters analyzed videos of all tasks. Raters analyzed videos independently from each other and were blinded to each other's results.

Video analysis

The video analysis method that was used in this study was described in detail earlier (Coenen et al., 2011). In short, beginning and ending frames of the task were selected from the video fragments by each rater. For lifting tasks, the start of a task was defined as the moment the load is clear from its surface, while the end of the task is the moment in which the end position of the load is reached. For pushing and pulling tasks, the task was defined as the period in which the worker is exposed to external forces at the hands due to resistance of the load. In addition, two intermediate frames, equally spaced in time between the beginning and end frame, were automatically selected to obtain four video frames. In these four video frames, a semi three-dimensional manikin was constructed consisting of nine segments (right foot, lower leg and upper leg; pelvis, trunk/head, two upper arms, two forearms/hands). This manikin allows for semi three-dimensional analysis of movements (ankle flexion/extension, knee flexion/extension, hip flexion/extension, trunk flexion/extension, trunk rotation, trunk lateral flexion, shoulder flexion/extension, shoulder abduction and elbow flexion/extension). Furthermore, the manikin can be scaled, rotated around its longitudinal axis (axial rotation) and translated horizontally and vertically along the video frame (Figure 7.1). Each rater made an optimal fit of the manikin to the four video frames for each of the 50 tasks by adjusting all segment orientations. Subsequently, for each task and rater, a cubic spline interpolation of the segment angles over the four key frames was executed to estimate body kinematics of the worker with a time resolution of 25 Hz. In case a MMH task lasted less than 2 s, only the first and the last frame instead of four video frames were used for cubic spline interpolation to avoid unrealistically high accelerations due to random errors in fitting the manikin. This interpolation method has been validated in a lab-based study before (Xu et al., 2010b). Based on total body mass and stature, individual segment masses and lengths, positions of the center of mass and inertia tensors were estimated using regression equations (Zatsiorsky, 2002). Hand forces were obtained from measured forces (at the time of video recording) in case of pushing and pulling, and from object weight (obtained at the time of video recording) and hand acceleration in case of lifting.

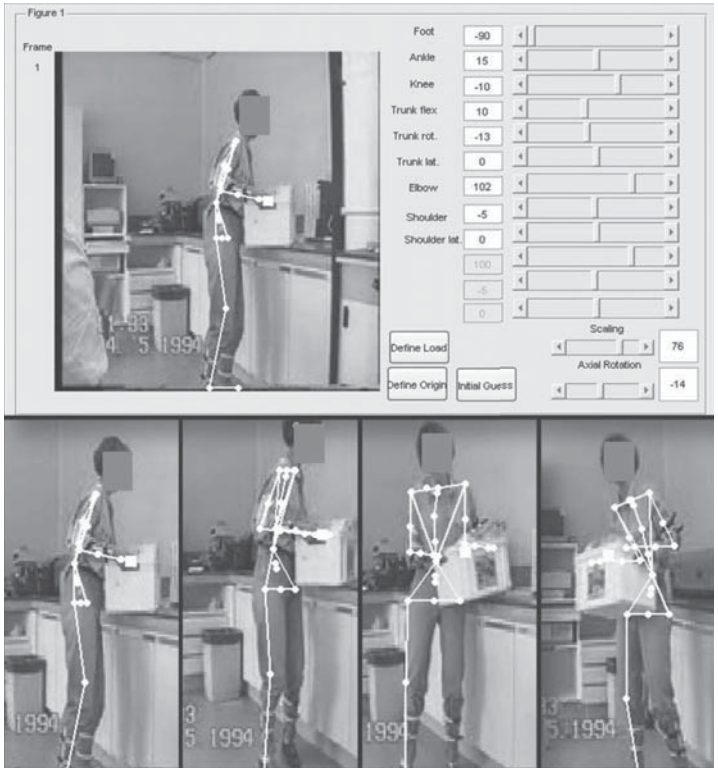

Figure 7.1 |Video analysis method. The graphical user interface depicting a three-dimensional manikin plotted onto a video frame is shown (upper part of the figure). In the lower part of the figure, a typical example of four key video frames of a field-based lifting task is shown that was analyzed by one of the observers.

A top-down inverse dynamics calculation using hand forces, segment kinematics (obtained from the interpolated manikin postures) and anthropometrics was performed to calculate dynamic moment components (derived from segment acceleration), static moment components (derived from gravitational forces on upper body segments and external forces at the hands) and total moments (static plus dynamic components) at the level of the L5S1 joint. For further analysis, the resultant moment (i.e., the resultant of the moments

around three axes) was considered. Both the moment at the instant of peak total moment and moments averaged over the entire task's time series were obtained. As horizontal load distances of the load with respect to the L5S1 joint is an important input variable for low-back load, horizontal low-back to load distance at the instant of peak moment was assessed. For further analyses, the abovementioned low-back load dose metrics and horizontal load distance and segment orientation angles at the instant of peak moment obtained from the interpolated manikin fit over the workers by each rater, were collected.

Data analysis

Intra-class correlation coefficients (ICCs) were calculated to assess the agreement among the five raters in the estimation of L5S1 peak and averaged moments (total moments; dynamic and static components of the moments), horizontal load distance and the segment angles. ICCs <0.40 were assumed poor, ICCs 0.40–0.75 were assumed good and ICCs >0.75 were assumed excellent (Fleiss, 1986). Furthermore, for the above-mentioned variables, standard deviations over the raters were calculated for each task while the median of these standard deviations over the 50 tasks was calculated to quantify inter-rater variability (Bao et al., 2009; Rothman & Greenland, 2005).

An additional analysis was performed in which inter-rater median standard deviations were assessed for lifting and for pushing/pulling tasks separately for peak and averaged total moments. This analysis was performed to test whether the variability among raters differed in lifting tasks compared to pushing/pulling tasks. Non-parametric Mann–Whitney-U tests were used to test for significant differences between lifting and pushing/pulling tasks assuming p-values <0.05 to be statistically significant.

RESULTS

Peak and mean moments across all tasks were on average 88.17 (15.83) Nm and 68.59 (11.39) Nm respectively. Furthermore, axial rotation across all tasks was on average 29 (31)° at the beginning of the tasks and changed on average 34 (67)° during the tasks.

ICCs were excellent for both peak (ICC = 0.92) and averaged (ICC = 0.91) L5S1 moments (Table 7.1). ICCs were substantially larger, but median inter-rater standard deviations were substantially larger as well for the static (ICC >0.90 and median standard deviation >8.2 Nm) compared to the dynamic (ICC <0.71 and median standard deviation <2.6 Nm) component of L5S1 moments, both with respect to peak (Table 7.1; Figure 7.2) and mean moments (Table 7.1; Figure 7.3). Concerning standard deviation of low-back moments, some occasional outliers for peak (>40 Nm) and mean moments (>30 Nm) were found (Figures 7.2 and 7.3).

ICCs of segment angles ranged from poor (trunk rotation and shoulder abduction), to good (trunk lateral flexion, shoulder flexion and elbow flexion) and excellent (trunk flexion; Table 7.1). Median standard deviations of the segment angles were low (<5°) for

the three trunk angles and for shoulder abduction and were higher (>14°) for elbow and shoulder flexion (Table 7.1). Resultant horizontal load distance with respect to the L5S1 joint showed small median standard deviation (0.08 m) and good ICCs. Non-parametric Mann–Whitney-U tests revealed no significant differences for median standard deviations of peak ($p = 0.64$) and mean moments ($p = 0.76$) between lifting and pushing/pulling tasks (Figure 7.4).

Table 7.1 | Absolute values (mean and standard deviation over 50 tasks after averaging over 5 observers) and inter-rater reliability estimates (intra-class correlation coefficient (ICC) and median over 50 tasks of the standard deviation over five observers) of low-back moments, and of segment angles and load distance at the instant of peak moment, obtained from the video analysis. Average values, standard deviations and median standard deviations are expressed in Nm for moments, in degrees for segment angles and in meters for load distance.

		Absolute Values		Inter-rater reliability	
Variable		Mean	Std.	ICC	Median Std.
Moments					
Peak moment	Total	88.17	15.83	0.92	8.80
	Static	79.96	12.92	0.93	8.85
	Dynamic	8.20	8.92	0.69	2.54
Mean moment	Total	68.59	11.39	0.91	8.31
	Static	63.65	11.22	0.91	8.63
	Dynamic	4.95	5.20	0.70	1.24
Segment angles					
Trunk flexion		13.87	2.60	0.91	3.58
Trunk rotation		0.14	5.07	0.26	4.89
Trunk lateral flexion		2.08	3.05	0.72	1.88
Elbow flexion right		72.35	10.81	0.63	16.22
Shoulder flexion right		26.33	10.11	0.61	14.49
Shoulder abduction right		4.83	10.36	0.33	4.25
Elbow flexion left		71.76	12.30	0.50	20.71
Shoulder flexion left		24.82	11.05	0.54	15.73
Shoulder abduction left		4.31	10.31	0.26	0.00
Load distance		0.43	0.16	0.63	0.08

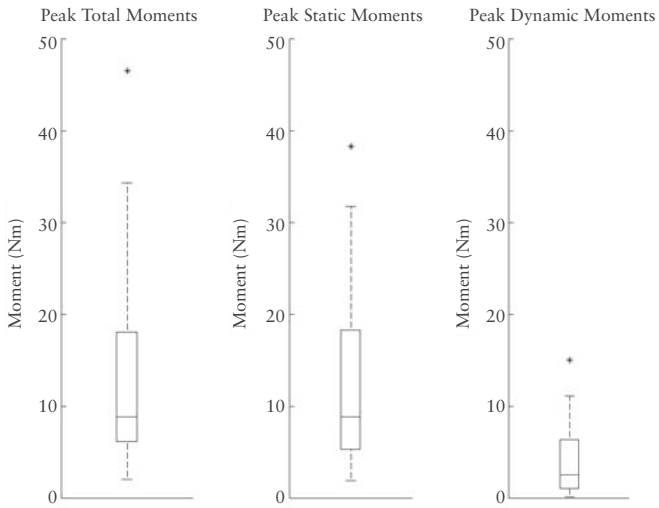

Figure 7.2 | Standard deviations across raters of all rated tasks concerning moments at the instant of peak of the total moment. The middle notch represents the median standard deviation, the box presents the standard deviations of the 25th percentile to the 75th percentile, whiskers represents the 5th to 95th percentile interval and asterisks represent outliers. Total moments (left plot), static component (middle plot) and dynamic component (right plot) of moments are shown. Values were calculated over all 50 tasks.

DISCUSSION

The aim of our study was to evaluate the inter-rater reliability of a video analysis method to estimate low-back load in work field situations. Our main focus was to assess low-back load dose estimates (i.e., low-back peak and mean moments) as these metrics are expected to provide more insight into low-back load than postural exposures (Wells et al., 2004), leading to stronger associations with LBP (Coenen et al., 2013b). Results show excellent ICCs for total low-back moment estimates. Median standard deviations assessing inter-rater variation were relatively low, i.e. about 10% of total moments. Inter-rater reliability was lower for dynamic components of the low-back moments compared to static components. The relatively low inter-rater reliability in dynamic moment components may partly be caused by the fact that inevitable random errors in positioning the manikin

are strongly magnified due to double differentiation of position and angle data (Xu et al., 2010b). However, as shown before (Coenen et al., 2011; van Dieën et al., 2010), dynamic components of the moments are only a small percentage of the total moment (i.e., about 10%; Table 7.1). Therefore, errors in dynamic components only contribute for a small part to errors in total moments. However, actual accelerations cannot be obtained from these data. The number of frames is a trade-off between the random errors in individual frames, the effect of which is increasingly magnified by differentiation when time intervals between frames are shorter, and the number of frames required to adequately cover the whole movement. It has been shown that using more than four frames does not improve the results when taking random errors in matching manikins to video frames into account (Xu et al., 2012). In the present study we observed that, as a result of the above-mentioned trade-off, for tasks lasting less than 2 s, using four frames resulted in unrealistically large accelerations. To avoid these unrealistically large accelerations, we decided to use the first and the last frame for interpolation instead of four video frames for tasks lasting less than 2 s. While Xu et al. (2012) showed that (random) errors increase by about 50% when taking 2 instead of 4 samples, we found in tasks with a duration less than 2 s that random errors caused unrealistic accelerations and a subsequent dramatic increase in inter-subject variation (up to over 100%). Regrettably, we could not check the validity of our approach to select 2 s as a threshold. Besides, in the study described by Xu and colleagues, only standardized tasks were studied in a laboratory situation, whereas we studied non-standardized field MMH tasks.

We found no significant differences in inter-rater variation of lifting tasks compared to pushing/pulling tasks for peak and averaged moments, suggesting that the current video analysis method is equally applicable to these three types of MMH tasks. As the tasks selected for our study were only a small proportion of all available tasks in the SMASH cohort, it can be argued that our selection may not be representative for the whole SMASH cohort or for MMH tasks in general. However, the tasks selected for our study were carefully chosen to cover a broad range of tasks from the original SMASH cohort with varying camera angles and occlusion of body segments. Therefore, the selection of workers and tasks used in the current study is considered representative for a broad range of workers, jobs and work settings. As an additional test, ICCs of the low-back loads within all subjects performing more than two tasks were assessed. These ICCs ranged from 0.68 to 0.99 for peak moments and from 0.42 to 0.99 for average moments. These results show that inter-rater agreement varied substantially across workers which is attributable to the variable quality and plane of video images across workers, as well as to the magnitude of the range of low-back loads within workers. While our findings may not be extrapolated to highly asymmetric or highly dynamic tasks, the high ICCS and low standard deviations in our low-back load estimates suggest that the proposed method is applicable for a broad range of tasks, both with and without asymmetry, variation in dynamics and load handled.

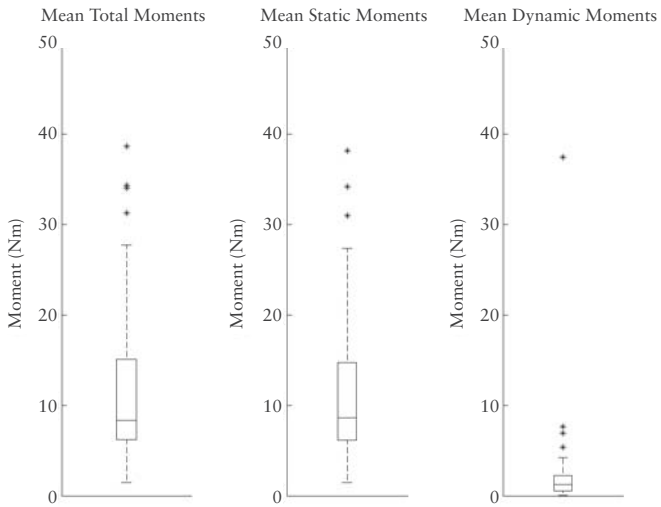

Figure 7.3 | Standard deviations across raters of all rated tasks concerning averaged moments. The middle notch represents the median standard deviation, the box presents the standard deviations of the 25th percentile to the 75th percentile, whiskers represents the 5th to 95th percentile interval and asterisks represent outliers. Total moments (left plot), static component (middle plot) and dynamic component (right plot) of moments are shown. Values were calculated over all 50 tasks.

Excellent inter-rater reliability was shown for trunk flexion angle; raters agreed well for trunk lateral flexion and elbow and shoulder flexion, however, agreement of trunk rotation and shoulder abduction was poor. In part, this may be due to less precise positioning of the manikin in the frontal and transverse plane relative to the sagittal plane. However, also median standard deviations showed varying inter-rater differences for segment angles. Since ICC is the ratio of the between task variance and the total variance (variance between tasks, variance between observers and random variance; Shrout & Fleiss, 1979), the ICC can be poor when the variance in observations is small (Bao et al., 2009). In our study, most raters estimated small movements outside the sagittal plane (e.g. trunk lateral bending, trunk rotation and shoulder abduction), leading to small variations in observations which can explain the poor ICCs for these segment orientations. For example, for shoulder abduction poor agreement was shown (ICCs of 0.33 and 0.26) that can be explained by rather small inter-rater standard deviations (4.25° and 0°; Table 7.1). In addition, trunk rotation and lateral flexion was rather small. However this was not due to little task asymmetry. Substantial asymmetry in the filming of tasks as well as axial rotation of the subjects during the tasks occurred as axial rotation across all tasks was on average 29 (31)° and changed on average 34 (67)°. Notably, however, workers mainly adapted to task asymmetry by whole body rotation rather than by adopting asymmetric postures.

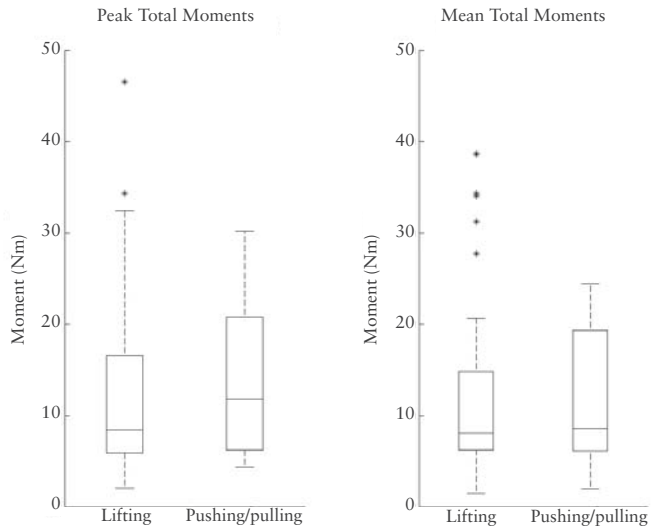

Figure 7.4 | Standard deviations across raters of all rated tasks concerning peak (left plot) and mean (right plot) moments calculated for lifting tasks only and for pushing and pulling tasks. In the figures, the middle notch represents the median standard deviation, the box presents the standard deviations of the 25th percentile to the 75th percentile, whiskers represents the 5th to 95th percentile interval and asterisks represent outliers. Values were calculated over all 50 tasks.

Despite relatively low inter-rater reliability of some postural variables, highly reliable low-back loads were found. A possible explanation is that not all postural variables contribute equally to the low-back load. For example, it is likely that the trunk flexion angle and the horizontal load distance with respect to the L5S1 segment contribute largely to the low-back moments whereas abduction of the shoulder contributes little to the low-back moment. In addition, an error in rating the shoulder angle can be compensated by a concomitant error in rating the elbow angle. This will then lead to a reliable load distance and consequent low-back load. This reasoning is supported by good inter-rater agreement for horizontal load distance of the load with respect to the L5S1 joint and of low-back moments, despite substantial errors in some of the posture variables. Furthermore, other postural variables (e.g., trunk flexion and trunk lateral flexion) do show highly reliable inter-rater reliability.

We did not compare our results to a gold standard as, regrettably, there is no gold standard in assessments of low-back load doses in field situations (Takala et al., 2010). Comparison of measurement tools described in other studies with respect to validity of outcomes is therefore difficult. However, in a lab-based validation study on the same video analysis method (Coenen et al., 2011) we found non-systematic, random errors for peak and mean low-back moments. The present study adds that between-rater differences are rather small (<10%), suggesting that the present video analysis method is a good method for low-back load assessments in field settings.

Although lab-based posture observation studies show comparable inter-observer agreement to the agreement reported here (Bao et al., 2009; Burt & Punnett, 1999), work-site postural observation methods, with and without the use of video recordings, have some drawbacks. They rely on crude categorical estimates, the magnitude of errors increases when joint angles become close to posture boundaries, outcomes heavily rely on the experience of the observer (Kociolek & Keir, 2010; Lowe, 2004; Spielholz et al., 2001), and observers seem to have difficulties to analyze more variables at once (Spielholz et al., 2001). Furthermore, agreement between raters is highly dependent on the number of categories used (Andrews et al., 2008). A postural variable categorized in a low number of categories is more likely to have a high inter-rater agreement, however, may lead to a loss of information (van Wyk et al., 2009). Eventually, large errors may result when using observations of working postures as input in biomechanical models estimating low-back load doses (de Looze et al., 1994b). Due to the reliable estimates of low-back moments and the on-line fitting of body orientations, the proposed video analysis method seems to be more appropriate to assess MMH tasks, especially when estimating low-back loads doses. In studies on comparable video coding systems, Xu et al. (2011) found, except for trunk lateral flexion, high ICCs (>0.75) for segment angles while Sullivan et al. (2002) found ICCs to be high as well for several low-back load metrics. These results are comparable to the ones reported here, however, both studies only reported on lab-tests, whereas we performed a study on field-based tasks.

Despite high inter-rater reliability and small variation among observers, relatively large errors can occur in some occasions. Such errors mainly occur in situations in which a part of the subject's body is occluded from view (e.g. when workers turn their back to the camera or when the view on the worker is, for example, occluded by another worker or by machinery). Although these substantial inter-rater differences occur in only a minor proportion of the tasks, such problems seem to be inevitable in field settings. The possible occurrence of these errors should therefore be noted when obtaining low-back load data from workers in field settings.

We used a relatively small number of raters who had substantial knowledge on kinesiology but no experience on working with low-back load assessment tools. External validity of the current video method can thus be questioned. However, our video analysis method is rather objective as it involves adjusting postures of the manikin to the posture of worker with continuous visual feedback of the manikin stick figure over the video frames. This procedure involves only minor subjective scoring, therefore, no major biases can be expected as a result of the selection of raters.

It has been reported in earlier studies that low-back loading is a risk factor for the occurrence of LBP (Marras et al., 2010; Norman et al., 1998). Both studies found significant differences in several low-back load metrics between workers with and without (risk of) LBP up to about 20%. The errors that we found between raters are substantially smaller than this percentage. Therefore, we expect only minor misclassifications in LBP risk groups due to inter rater variability using the proposed video analysis method.

CONCLUSIONS

The current study shows that the proposed video analysis method is reliable when used by different raters, which makes it applicable in epidemiological studies or ergonomic practice for low-back load dose assessment. Inter-rater reliability for low-back moments is high, while the agreement for rating of the most important segment angles is reasonable. Errors are small enough to limit the likeliness of misclassification in LBP risk groups. Although occasional substantial errors can be made when assessing MMH tasks, this study shows good overall agreement among raters.