Work 40 (2011) 445–457
DOI 10.3233/WOR-2011-1256
IOS Press

Accuracy of visual estimation in classifying effort during a lifting task

Darrell W. Schapmire[a,*], James D. St. James[b], Robert Townsend[c] and Larry Feeler[d]
[a]*X-RTS Software Products and Testing Devices, Hopedale, IL, USA*
[b]*Millikin University, Decatur, IL, USA*
[c]*WCS Occupational Rehabilitation Sports Medicine, Palos Heights, IL, USA*
[d]*WorkSTEPS, Austin, TX, USA*

Received 7 December 2009
Accepted 14 March 2010

Abstract. *Objective:* The objective was to determine if visual estimation of effort (VEE) during lifting tasks is accurate in classifying relative levels of exertion or distinguishing between incomplete lifts that may be potentially unsafe and incomplete lifts of "actors" feigning weakness.
Participants: A convenience sample of 117 health professionals and lay subjects participated in the study.
Methods: Four actors were videoed performing four complete dynamic lifts (sets of five repetitions) of varying levels of exertion (relative to subjects' physical maximum). Subjects viewed the videoed performances, presented in no apparent order, attempting to properly classify the lifting tasks. For the four levels of exertion, participants were to judge if the lifts were 25%, 50%, 75% and 100% of each actor's maximum lifting capacity and to distinguish between an incomplete (failed) lift of 110% of maximum and a feigned failure of a lift of 25% of maximum.
Results: Accuracy for in classifying all lifting activities was marginally higher than chance. There were no differences in the accuracy of health professionals or lay subjects.
Conclusion: The VEE does not accurately classify relative levels of exertion or distinguish between incomplete feigned effort lifts and lifts that are potentially too heavy to safely lift.

Keywords: Psychophysical, kinesiophysical, lifting evaluation, functional capacity evaluation (FCE)

1. Background

1.1. Application of the VEE in psychophysical and kinesiophysical methodology

Lifting is an essential function of many occupations. There have been two methods proposed to use visual estimation of effort (VEE) for the purported purpose of classifying the relative level of effort or exertion of a lifting task and to determine if an individual is giving a maximum voluntary effort during the assessment. Those methods are known as the "psychophysical" approach and the "kinesiophysical" approach.

Early studies on the psychophysical method were based on physical performance data collected on volunteers for study, prospective employees and incumbent employees [4,5,10,25,36–38]. The psychophysical method had many variants, with no single testing variant being universally adapted to identify maximum, safe levels of exertion during a lifting task. Each version of the psychophysical approach used "operational definitions" (essentially, observations of the test administrator) and input from the subject to determine when the maximum safe level of lifting was attained. However, none of the aforementioned studies provide specific sets of operational definitions.

*Address for correspondence: Darrell Schapmire, MS, X-RTS Software Products, Inc., P.O. Box 171, Hopedale, IL 61747, USA. Tel.: +1 309 449 5483; E-mail: ds@xrts.com.

Table 1
Observation criteria for level of effort [18]

	Light	Moderate	Heavy
Muscle Recruitment	No accessory muscle; recruitment; prime movers only, (quadriceps, trunk stabilizers, biceps, hand grip)	Recruitment of accessory muscles, neck flexors, upper trapezius, deltoids	Pronounced recruitment of neck flexors, trapezius, deltoids, rhomboids
Body Mechanics	Safe	Safe	Safe
Base of Support	Natural posture	Stable base	Very wide, solid base
Posture	Upright posture	Beginning of counterbalance in extension	Marked, increased counterbalancing
Control and Safety	Easy movement patterns	Smooth movements: Increase time of lift test	Uses momentum in a controlled manner, increase time of lift test

In the kinesiophysical methodology, specific operational definitions for identifying exertion at the "light," "moderate" and "heavy" levels are used [18,30]. These operational definitions are shown in Table 1. A variant of the kinesiophysical approach described above has been proposed in another study [20]. In this methodology, "Criteria for determining a maximum physical effort include visible changes in trunk and limb alignment, visible evidence of muscle fatigue, contraction of accessory muscles, and body movements that the therapist believes indicate compensation for fatiguing muscles (p. 999)" However, no specifics were reported with regard to "how much" change in trunk alignment is permissible, how to physically measure changes in trunk alignment, how to visually distinguish between actual or feigned muscular fatigue, which accessory muscles were observed, or any guidelines that would help distinguish between movement and movement that was not genuinely compensatory.

The primary difference between the psychophysical and kinesiophysical methodology is in the termination points for "maximum lifting." In the former, a physical maximum is thought to have been attained if the evaluator believes that changes in body compensatory mechanics indicate that heavier lifting would be unsafe, or if the subject indicates that a heavier workload would be unsafe. By contrast, in the kinesiophysical approach, the evaluator controls the termination point for maximum lifting. The substantial similarity between these testing approaches is that if the subject terminates the lifting activity at a level believed by the evaluator to be less than a maximum effort, the subject is generally believed to have been uncooperative, self-limiting, or performing at a less-than-maximal level. There are essentially no built-in cross-validation methods to ensure the internal validity of the lifting data. The opinion of the evaluator is considered to be the final authority.

By and large, both the psychophysical and kinesiophysical methodologies have been promoted as commercial testing methods on the basis of studies of reliability that examine the consistency of ratings (degree of agreement) either based on agreement between raters (inter-rater reliability), agreement of repeated ratings by the same rater (intra-rater reliability), test-retest reliability, or agreement across protocols [2,8,11–14,16,17,22,23,26–30,33,29–42,44]. Unfortunately, reliability does not address whether the ratings were accurate. Accuracy is the percentage of correct classification. If two raters make the same errors in classification of effort, their reliability would still be very high.

Three previous controlled studies [3,19,21] reported sensitivity and specificity in the classification of overall effort during commercially-available functional assessments [3,19,21]. The sensitivity and specificity reported in those studies, however, included observers' classification of effort during dynamic lifting capacities, hand strength, various non-material-handling activities and static strength testing. However, no controlled studies regarding the accuracy of the commercially-available FCE protocols *focusing strictly on the accuracy of relative levels of exertion during lifting activity*, have been identified by the authors.

In a rare reference to the accuracy of the VEE from the body of the reliability studies, one study reported that, "-maximal' performances were correctly rated in 46% to 53% (healthy subjects) and in 5% to 7% (patients with chronic nonspecific low back pain) of the cases (p. E40)" [30]. The various reliability results in the study, however, ranged from 0.50 to 0.92. Thus, it should be understood that "high reliability" does not necessarily indicate "high accuracy" in identifying maximum levels of lifting capacity.

1.2. Controlled studies addressing accuracy

The psychophysical approach has been assessed for its accuracy in identifying maximal and sub-maximal effort during an FCE in controlled studies [3,19,21]. However, these studies did not report specific criteria for classifying exertion and effort during a lifting task.

Table 2
Actor demographics and reasons for determining "Max"

Subject	Age	Height	Body weight	Max Lift	Reason for termination at heaviest level of lifting	Relevant medical history
Female 1	42	1.67 m	58.96 kg	23.26 kg	Lifting additional weight caused pain in left gluteals; terminated by actor with relevant medical history described in next column	History of lumbar disc herniation, intermittent treatment by physical therapist
Female 2	34	1.70 m	70.76 kg	27.80 kg	Unable to lift more weight to waist level (biceps as limiting factor)	None
Male 1	36	1.85 m	106.59 kg	79.96 kg	Unable to lift more weight to waist level (biceps as limiting factor)	None
Male 2	58	1.77 m	97.52 kg	48.21 kg	Right SI joint pain; terminated by actor with relevant medical history described in next column	History of laminectomy/foramenotomy, history of short-term (3 week) period of disability and treatment for SI joint dysfunction one month prior to experiment

There has been one reliability study using a kinesiophysical assessment method that reported 100% accuracy and a *kappa* coefficient of 1.0 – perfect agreement – between evaluators in ratings of effort of lifting as "light," "moderate" or "heavy" workloads, using the VEE [18]. The first author of that study rated each lift performed during the videotaped lifting sessions of different "actors," according to a set of operational definitions. The raters, trained in the use of these operational definitions in clinical practice, were scored in accordance with their agreement with those ratings. In effect, this reduces the data to an inter-rater reliability study in which one person's judgment becomes the standard against which all others are compared, with no absolute guarantee that the expert is correct in all cases. In calculating the *kappa* coefficient, lifting activities judged by the first author to be at the "moderate" level were omitted from the analysis. Furthermore, ratings of either "heavy" or "moderate" were counted as "correct" for the lifts classified as "heavy" by the first author. Lastly, ratings of "light" or "moderate" were counted as accurate for lifts classified as "light" by the first author. Note that with this scoring system, there would be 100% accuracy and perfect kappa coefficient of 1.0 if every rater gave a rating of moderate.

1.3. Purpose and general areas of investigation of this study

Given the scarcity of studies that speak to the accuracy of the VEE during a lifting task, the purpose of this study was to explore the issue of accuracy. Specifically, it was the intent of the authors to determine the accuracy of VEE with regard to the ability to correctly identify relative levels of exertion (lifting 25%, 50%, 75% and 100% of a "safe maximum lift").

Also investigated was the accuracy in distinguishing between lifts that were incomplete because the actors were lifting beyond their measured safe maximum lift (a failed lift of 110% of maximum) and incomplete lifts of 25% of maximum, in which the actor pretended to fail to lift the weight (feigned weakness).

2. Methods

The experiment was conducted under the auspices of the Millikin University (Decatur, IL) Institutional Review Board. Four actors performed the lifting activities in this experiment. Demographic information pertaining to the actors is found in Table 2. A psychophysical maximum for a bilateral lift to waist level, initiated with the knuckles approximately 0.31 m above floor level, was determined for each actor. The lifting activities were terminated for reasons listed in Table 2. To replicate the actual progression of workloads that might be lifted during an FCE (reference being made to the first four lifting activities in the list below), each actor then performed the sequence of activities in the order listed below:

1. Five repetitions at 25% of the actor's physical maximum.
2. Five repetitions at 50% of the actor's physical maximum.
3. Five repetitions at 75% of the actor's physical maximum.
4. Five repetitions at 100% of the actor's physical maximum.
5. A single incomplete lift through approximately half of the distance to the waist at 110% of the actor's physical maximum.

6. A single incomplete lift through approximately half of the distance to the waist at 25% of the actor's physical maximum.

After completing each lift in the set of five repetitions, the actors returned to an upright posture prior to initiating the subsequent lift. During the incomplete lifts, the actors were instructed to lift the loads through approximately 50% of the lifting range of motion prior to returning the workload to its starting position. With the exception of the incomplete lift of 25% of maximum, the actors made no attempt to misrepresent the level of difficulty of any of the lifting activities. Actors were told to perform the lifting activities at the pace of their choice. A video camera filmed each actor from the left side as the various lifting activities were performed. In order to give a clearer view of the right upper extremity, but to do so without excessively obscuring the subjects' view of the back, the actor's frontal plane was at 10 degrees left rotation, relative to the camera. See Fig. 1. The camera was stationary at a fixed point throughout the video taping. The height of the shortest subject, relative to the amount of vertical space on the viewing area for the media player, was 78%. The tallest subject's height for the same comparison was 89%. The sound was muted.

The videos of the lifting performances were then scrambled into a sequence of 24 different scenes, with the sequencing meeting these criteria:

1. The first scene was not 100% of any actor's physical maximum.
2. No successive scenes depict the same actor.
3. No successive scenes depict lifts of loads that are at the same relative level of exertion.
4. No successive scenes depict incomplete lifts.
5. Half of the incomplete lifts at 25% and 110% of the respective actor's maximum were in the first 12 scenes, and the remaining were in the second half of the sequence.

A second sequence of the scenes was configured by reversing the order of the 24 scenes.

The total running time for the video was 16 minutes 45 seconds, including 1 minute 5 seconds for re-stating the written instructions, one practice trial, and 15 seconds pause between each individual scene. The text on the monitor between scenes informed the viewer as to whether the subsequent scene would be a set of five repetitions of the same lift or a single incomplete lift of either 25% or 110% of the subject's maximum lifting capacity. Subjects viewed the video on a personal computer or laptop. Approximately half the subjects viewed the first sequence and the remainder viewed the second sequence. Data collection sheets were either faxed or mailed to DS.

Fig. 1. View of actor participating in study, shown returning to a standing posture after lifting a workload.

The subjects for this study were a convenience sample. Prior to viewing the videotaped performances, the subjects were given a set of written instructions. All subjects taking part in this study who had conducted FCEs had received training in at least one FCE training protocol, all of which employed a VEE. These subjects were instructed to use the operational definitions of their choice, based on their training, for classifying each scene in the video. The lay and the medical and allied health care professional subjects had no training in FCE methodology. They were instructed "to make the best judgments you can" when classifying the activities in the videos. All subjects were instructed to view the 24 scenes in the experiment without stopping or replaying any portion of the video. Subjects were informed in the written instructions and in the video frames between the various scenes that the repetitive lifts would be equal to 25%, 50%, 75% or 100% of the actors' maximum lifting capacity. They were also informed that they would observe incomplete lifts of 25% and of 110% of each actor's maximum lifting capacity. The written instructions specifically instructed the subjects to "make no assumptions" as to how many

Table 3
Education (lay subjects *v.* all health care providers)

	Non-degreed	Associate degree	Bachelor degree	Master degree	DPt or PhD	Medical doctor
Lay Subjects	12 (60%)	0	8 (40%)	0	0	0
All Health Care Professionals	6 (6.2%)	6 (6.2%)	33 (34.0%)	32 (33%)	14 (14.4%)	6 (6.2%)

times each actor would or would not perform any specific lifting task. The subjects circled their estimates of relative levels of effort on a data collection sheet.

3. Results

There were 63 subjects who viewed Sequence 1 and 54 subjects who viewed Sequence 2. The two sequences were combined for the total of 117 subjects (56 males and 61 females). There were three groups of subjects in this study:

1. 20 lay subjects, having no training or experience in physical testing;
2. 59 health or allied professionals having no training or experience in administering lifting evaluations (23 physical therapists, 13 physical therapist assistants, 6 medical doctors, 2 nurses, 2 physical therapy techs with bachelor degrees, 8 physical therapy technicians, 5 miscellaneous); and
3. 38 health professionals having training in at least one commercial FCE testing methodology and also had experience in performing lifting assessments (33 physical therapists, 3 occupational therapists, 1 athletic trainer, 1 exercise physiologist).

The mean age for the subjects in this study was 40.8 years. For subjects with training and experience in administering FCEs, the mean number of tests administered, per the estimation of the each subject, was 259.41 (SD 332.49, Range 3–1,000). The mean number of years experience administering FCEs was 8.31 (SD 7.40). The mean number of years professional experience was 12.83 (SD 9.00) for subjects performing FCEs as compared to 13.09 years (SD 9.84) for allied health care professionals who were never trained and did not perform FCEs. Education (highest degree) for all subjects and total years of professional experience are shown in Table 3.

3.1. Accuracy in classification of relative levels of effort

Table 4 reports the accuracy for estimation or the relative levels of exertion during the lifting of workloads

Table 4
Accuracy of visual estimations (percent correct classification) for each relative level of effort

	25%	50%	75%	100%
All subjects	55.4	37.9	35.4	33.2

equal to 25%, 50%, 75% and 100% for all subjects. Table 5 reports the accuracy per group membership. For percent correct classification, the mean difference approached significance, $F(2, 114) = 2.62$, $p = 0.077$, though it accounted for only 4.4% of the variability in ratings. It is noted that the level of accuracy for lay subjects and trained, experienced evaluators was nearly identical. As shown in Table 5 the actual differences were small, and the difference in percent correct was near zero for the comparison of trained therapists currently performing FCEs and persons with no training or experience in the visual estimation of effort. For absolute error, there was no statistical difference, $F < 1$. The difference between the group means accounted for only 1.6% of the variability in the percentage of errors. The absolute error is the number of percentage points an estimation of effort differed from the correct response. In other words, if the scene depicted lifting at the 50% level and the estimation was 50%, the absolute error in percent was 0%. But if the estimation was 75% when the activity was at the 50% level, the absolute error would be 25%. The data in Table 5 show average absolute error ranges between 18.5% and 20.0% for the three groups. In other words, the average error for each group approaches one entire relative level of exertion in the four levels assessed in this study.

3.2. Correlation between accuracy and demographic factors

While the overall accuracy was low, a question of major concern for this research is whether the training, education, and experience of the therapist affect accuracy of their judgments. We performed various analyses to examine this issue.

Correlations were performed to examine the impact of various measures of experience. These are reported in Table 6. For each measure, data were missing for some subjects. Neither the subjects' ages, their number

Table 5
Percent correct in classifying the four levels of exertion, per group status

	Percent correct classification	Absolute error
Lay subjects	43.8 (SD = 12.0)	18.5 (SD = 6.0)
Untrained health professionals	37.8 (SD = 12.0)	20.0 (SD = 5.2)
Trained and experienced health professionals	43.1 (SD = 14.9)	18.7 (SD = 6.1)

Table 6
Correlations with various measures of experience

	Percent correct
Age	$r(113) = 0.015$, $p = 0.873$
Years experience	$r(72) = 0.069$, $p = 0.560$
Years testing	$r(36) = 0.063$, $p = 0.705$
Number of tests	$r(37) = 0.021$, $p = 0.898$

Table 7
Accuracy for identifying all four relative levels of exertion as a function of educational level for all health care professionals

	n	Percent correct
Non-degreed	6	34.4
Associates degree	6	31.3
Bachelors degree	33	41.7
Masters degree	32	41.1
Doctorate or medical doctor	20	40.0

of years of experience as a therapist, their number of years of experience performing FCEs, nor the approximate number of FCEs each had performed was significantly related to the percentage of correct classification. The amount of experience was unrelated to either the percentage of correct classifications or the mean error in classification.

Table 7 reports the percent correct classification for the four levels of education represented in our sample Percent correct classification did not differ significantly across educational levels, F < 1.

3.3. *Identification of feigned weakness and of potentially unsafe lifting during incomplete lifts*

In actual functional assessments, it is not uncommon for clients to demonstrate the inability to complete a lift. In such cases, it is necessary to answer this question: Is the failure due to physical limitation or it is due to a factor other than physical limitation? In Table 8, we examine the claims in regard to detecting feigned weakness in lifting by assessing the subjects' ability to distinguish between the scenes depicting incomplete lifts of 25% of maximum and 110% of maximum lifting capacity. The incomplete lifts of 25% depict feigned weakness. [Note to Reviewer: A sentence in the original manuscript made reference to "no deceptive intent," which was an oversight and pertained only to the repetitive lifts. That has been corrected/clarified. Our intent

Table 8
Mean percent correct classification for classifying incomplete lifts per group membership

	Sensitivity	Specificity
Lay subjects	54.8 (*SD* = 25.8)	71.4 (*SD* = 27.7)
Untrained medical professionals	62.3 (*SD* = 22.2)	71.9 (*SD* = 18.3)
Trained and experienced medical professionals	64.5 (*SD* = 22.3)	66.2 (*SD* = 23.7)

with the "no deceptive intent" descriptor was to ensure the reader that during the repetitive lifts, no attempt was made to misrepresent the level of difficulty. In abstract in the original manuscript reference is made to "feigned weakness" during the 25% lift.] The incomplete lifts of 110% of maximum lifting capacity depict the actors initiating a lift that would be potentially unsafe to complete.

For the purposes of this analysis, *sensitivity* involves correctly detecting feigned weakness. It is measured as the percent of trials in which subjects correctly indicate that the actor is feigning weakness. *Specificity* involves correctly detecting a failure of a supra-maximal lift. It is measured as the percent of trials in which subjects correctly indicate that the actor is failing to lift at a supra-maximal level. Two subjects each had two ratings missing for the incomplete lifts. Their data were discarded. None of the differences among groups is significant. For sensitivity, $F(2, 112) = 1.49$, $p = 0.229$. For specificity, $F < 1$.

4. Discussion

4.1. *Implications, professional and ethical considerations*

These findings clearly indicate VEE is not accurate in classifying relative levels of exertion, nor is it accurate in classifying relative levels of exertion or in determining if incomplete lifts were incomplete because the workload was potentially unsafe to lift or because the lifter was potentially feigning weakness. These results have implications for all VEE protocols, whether such protocols are used for the purpose of testing prospective employees or testing insurance claimants in an FCE.

For post-offer testing, a VEE that is enhanced by a process that has been refined, contains internal cross-validation and is legally defensible for post-offer testing purposes has been developed [9]. This protocol does not rely solely on the client's subjective report of "difficulty" to determine an end point for lifting activities because job applicants are potentially motivated to perform lifting tasks that may exceed a safe level of exertion. Furthermore, the protocol does not rely solely on the discretion of the evaluator to terminate lifting activities. Rather, there are built-in cross-validation measures to ensure the safety of the testing protocol and the integrity of the data. For assessing insurance claimants during an FCE, a test-retest and distraction-based protocol has been suggested as a viable alternative to the "standard" VEE protocols [43].

VEE is the most widely-used method for assessing exertion and cooperation. Its value as a testing method has been accepted, apparently on the basis of reliability studies. Therefore, it is incumbent on the authors to discuss the many reasons the VEE lacks sufficient accuracy to be considered as a "scientific" method of assessment.

On a professional and ethical level, the authors ask this question: Which type of legal machination would be most likely to result in the delivery of timely, appropriate care and better outcomes: a system which relies on the opinions and impressions of health care professionals who testify as expert witnesses, or a system which requires the expert witness to produce evidence that can be shown to have a scientific basis?

4.2. *Effects of special training, education and experience on accuracy*

Special training, educational level, years of experience and number of tests conducted have no statistical relationship to accuracy in this study. There were no differences across the three groups in this study with regard to the level of accuracy in using the VEE to identify lifting activities when actors lifted workloads equal to 25%, 50%, 75% and 100% of their maximum capacities. The level of accuracy was marginally higher than chance. Furthermore, the accuracy in classifying the feigned incomplete lifts of 25% of maximum and actual incomplete lifts of 110% of maximum was essentially the same for all groups – somewhat better than chance. .

4.3. *Conceptual flaws of VEE and the weakness of "reliability" statistics*

4.3.1. *Basic conceptual flaw – use of non-numeric descriptors*

The application of any set of operational definitions which ascribe various characteristics to different levels of exertion or cooperation is conceptually flawed from the outset. First, the ability to lift weight represents data that exists on a continuum that has many potential intermediate points – not just "light," and "heavy" with one intermediate level of difficulty between the extremes as described in two previous studies [18,30]. Words alone are incapable of adequately describing or classifying exertion. Second, the use of operational definitions becomes even less precise for any given lifting performance when there are believed to be characteristics present that would be associated with multiple levels of exertion. In other words, it is possible an evaluator would perceive "pronounced recruitment," a characteristic of "heavy" lifting, while simultaneously observing "onset of counter-balancing," said to be a characteristic of "moderate" lifting, according to two studies [18,30]. How are such instances interpreted, and how can such occurrences be interpreted in a standardized manner? No specific scoring or weighting mechanism is suggested for these "mixed results" scenarios – which surely occur in actual clinical practice.

Non-numeric operational definitions also lack precision and objectivity. For example, it is questioned how an observer might objectively distinguish between the "easy movement patterns" said to be present during "light" lifting and the "smooth movements" said to be present during "moderate" lifting, as proposed in two studies [18,30]. The authors also question how one might objectively determine if there is "recruitment of accessory muscles during "light" lifting and the "pronounced recruitment" that is supposedly observed during "heavy" lifting. Likewise, operational definitions such as "changes in trunk alignment and limb alignment, visible evidence of muscle fatigue, contraction of accessory muscles," as described in another study [20] are similarly vague. It is submitted that impressions of such vague criteria cannot be standardized between observers. These difficulties might be overlooked if not for the other substantial problems with such assessments.

4.3.2. *Inherent weakness of correlation statistics*

Studies of inter-tester reliability are concerned with the degree to which different observers of the same act

Fig. 2. Sam Tsang preparing to lift 206 kg, courtesy of Sam Tsang (athlete) and Dave Draper (site owner), http://www.davedraper.com (IronOnLine).

give consistent ratings – do people agree with each other? These studies are reported as *kappa* coefficients or other correlation statistics. Other studies investigating intra-rater reliabilities, based on similar measures, are concerned with the consistency of ratings made by the same observers at different times – do people agree with themselves?

4.4. *Use of "base of support" and disregard for effects of lever arm lengths*

In two studies of the kinesiophysical method, the authors suggest the use of a "stable base" was said to be associated with "moderate" lifting [18,30]. A "wider, very solid base" was associated with "heavy" lifting. In addition to a lack of precision in the operational definitions, it appears that personal preference may largely account for foot placement. Figure 1 clearly indicates individual preferences for foot placement when lifting heavy loads may not, in fact, be "wide." Figure 2 depicts another individual lifting a comparable workload with very wide foot placement. Although this is anecdotal evidence, it is reasonable to believe that foot placement would also be a personal preference or habit for non-weight lifters.

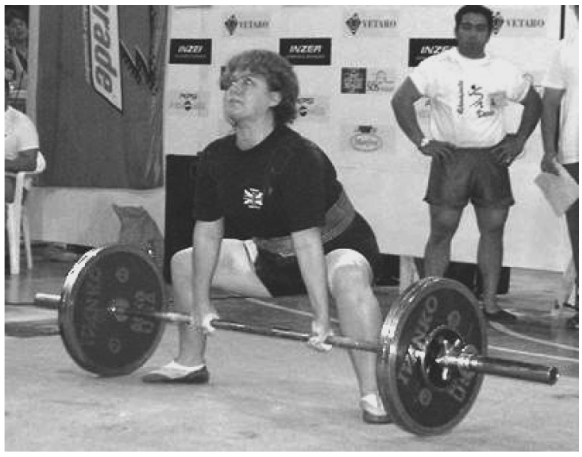

Fig. 3. Catherine Wass preparing to lift 170 kg, courtesy of Catherine Wass (athlete) and Stuart Hamilton (site owner), http://www.hamiltonsfitness.co.uk (Hamiltons Health and Fitness Ltd).

4.5. *Limitations of vision and other confounding variables*

4.5.1. *Limitations of foveal vision*

It is not readily apparent on every-day observation, but humans actually see the clearly in a narrow range of our central visual field. This is our *foveal vision*. It is approximately 2 degrees of angle in width and height. At 3 m, foveal vision would encompass an area of slightly more than 10.1 cm in diameter – a little smaller than a typical person's face at that distance. At shorter distances similar to the typical distance between evaluator and subject, the area of foveal vision is even smaller. Visual acuity (eyesight) deteriorates drastically as images move outside the fovea into the peripheral retina. A person with 20/20 vision straight ahead typically has acuity of 20/200 or worse only *five* degrees outside the fovea [6].

The narrow area of foveal vision can easily be demonstrated. Look at the X in the middle of the line of letters below, and then try to read the letters off to the side without moving your eyes. At normal reading distance, most people can recognize at most about 3 letters to each side of the X.

KDFLRPxGJRQSN

This illustration should be sufficient to point out that simultaneous visual observation of the shoulders, upper extremities, low back and lower extremities is not possible within our foveal vision, given the standard practice of an evaluator standing only a short distance from a client when using a visual estimation of effort

approach. Furthermore, since the subjects in this study watched videos of lifting activity, they had an advantage in terms of visual acuity they would not enjoy when attempting to observe multiple body parts simultaneously during an actual test [15].

Why do we not ordinarily notice the poor acuity outside of foveal vision? A major reason is that we frequently move our eyes, thus seeing quite a bit of the scene clearly over a time of a few seconds – but very little of it during any one fixation. We also see *movement* throughout peripheral vision [1]. The reader can demonstrate that movement is easily seen in extreme peripheral vision by holding a hand off to one side and waving your fingers. That something is moving was very easily appreciated, though it is not possible to accurately recognize details of what is seen. It is also the case that we see color fairly well out to about 50 degrees from straight head, and patches of color moving surely tell us quite a bit, especially if they are *familiar* images – for example, a family member seen out of the corner of the eye may be easily recognized.

We raise this issue of visual acuity outside the fovea in the context of the operational definitions of any visual estimation methodologies. Operational definitions relative to muscles in the lower extremities, upper extremities, the posterior aspect of the torso and the neck imply that the observer will simultaneously evaluate the level of recruitment of both prime movers (quadriceps, trunk stabilizers, biceps and hand grips) for light lifting and the accessory muscles (neck flexors, trapezius, deltoids, rhomboids) recruited during heavier lifting. Even at a distance of 3 m from the person lifting, it would be impossible to simultaneously inspect contractions of all of the muscles upon which the operational definition is based. In fact, the task would be even more difficult from a closer range because the area of the "window" of foveal vision becomes smaller as the distance between the eye and the object becomes shorter. Such observations are impressions, not objective fact.

4.6. *Practical challenge in "visually observing" muscular contractions*

The operational definitions in previous studies specifically reference recruitment of neck flexors, upper trapezius, and deltoids as a characteristic of "moderate" work and pronounced recruitment of neck flexors, trapezius, deltoids and rhomboids as a characteristic of "heavy" work [18,30]. Even if persons presenting for an FCE wear athletic shorts and short sleeved shirts, of all the aforementioned muscle groups, only the neck flexors are fully visible. Certainly, the evaluator's ability to observe the upper trapezius, deltoids, and even the prime movers (quadriceps, trunk stabilizers and biceps) is compromised at best. Furthermore the rhomboids, which lie under the trapezius, cannot be viewed without dissection of the back.

4.6.1. *Purported use of visual observations of "muscular contraction" as an index of effort*

Observations relative to muscular contraction are flawed on yet another level. Contraction of a group of muscles may *imply* that work is being performed – but cannot be associated with any specific level of work. It is, after all, possible to contract the biceps (or any other muscle group, for that matter) and thereby increase its apparent size – during an isometric contraction during which no lifting is being performed.

It is tempting to say that an evaluator is able to deduce muscle contraction on the basis of movements that are observed. However, the authors question how an evaluator would distinguish between movements that occur as the result of contraction of any particular group of muscles during normal, unimpaired, pain-free movement, as opposed to movements that occur as the result of compensatory movement strategies that might occur as the result of pain or orthopedic dysfunction. Unless such movements are obviously and grossly unusual, visual observations are scarcely more than descriptive guesswork – and even if correct are not objective. Observations relative to "muscular contraction" seem to have a limited role in objectively assessing exertion or cooperation.

4.6.2. *Inattentional blindness*

The use of operational definitions to classify effort and cooperation necessarily involves an attempt to visually monitor and be cognizant of multiple physical characteristics during a dynamic activity. It has been conclusively demonstrated, however, that subjects tasked with making a single, specific observation during a series of dynamic events will fail to identify other significant events that occur simultaneous to the specific event that is being observed [34]. This phenomenon is called inattentional blindness, which is illustrated in Simons' and Chabris' now-classic experiment in cognitive perception: http://www.theinvisiblegorilla.com/videos.html.

Table 9
Time, in seconds, to complete five lifts for each relative level of exertion

Subject	25%	50%	75%	100%
Female 1	34	33	31	38
Female 2	26	26	27	28
Male 1	28	24	28	30
Male 2	31	30	31	29

4.7. *Use of "momentum," "time to lift," and heart rate as adjuncts*

4.7.1. *"Increased time of lift" and "use of momentum" as observational criteria*

Two studies mention "increased time of lift" as a characteristic for "moderate" and "heavy" lifting [18, 30]. However, no mention is made in theses studies of the use of a stop watch or clock to document these changes. Therefore, the authors question how "time" was objectively incorporated into the analysis of the results. Furthermore, the operational definition for "heavy" lifting includes a reference to the use of "momentum" to complete the lift. Momentum is defined as "mass x velocity." The authors point out that unless an electronic system of evaluation is used to measure velocity in distance or degrees per unit of time, it is impossible to make accurate and objective assessments of the use of momentum. Certainly, the subjects in the studies in question had no firsthand knowledge of the amount of weight lifted by the actors. Therefore, any determination of perceived "use of momentum" by the subjects was speculation, not an objective observation. In addition, the simultaneous presence of "increased time of lift test" and the use of greater momentum may be mathematically contradictory, specifically depending on the time to perform the lift and the amount of weight lifted. Lastly, on the experiential side, the results reported in Table 9 indicate that "time of lift" was not a discriminating factor in differentiating levels of exertion for the four actors in this study. Rather, the amount of time involved in performing a set of lifts appears to be a personal habit or preference of the actors.

4.7.2. *Heart rate increases as an index of effort*

Physical work may increase heart rate, in the absence of the effects of medication. Since heart rate monitors are relatively inexpensive, easy to use and are presumably accurate, heart rate is sometimes used in lifting assessments to gauge the relative difficulty and, in some instances, used as an index of effort. One of the most common methods suggested to monitor exertion and cooperation during a lifting assessment is to choose an *arbitrary* cutoff with respect to an *age-predicted* maximum heart rate. But how is a maximum heart rate estimated, and are such estimates accurate?

One exhaustive review of textbooks, dissertations and studies published between 1957 and 2000 on the subject of maximum heart rate (HR) determined that the standard error of estimate for the widely-used formula "Max HR = 220 – Age" is approximately 10 beats per minute (BPM) for healthy subjects [31]. Therefore, to use this formula and be 95% certain of the prediction, the prediction must be expressed in a *range* of 40 beats per minute. Thus, the age-predicted maximum heart rate for a forty-year old person is normally assumed to be 180 BPM. But in reality, the maximum heart rate for a person of 40 years is at least 160 BPM, but no higher than 200 BPM. Therefore, using an arbitrary cutoff of, say, "70% of max predicted HR," could conceivably be as low as 112 BPM, as high as 140 BPM – or anywhere within that range. The only way to determine maximum HR is to conduct a maximum graded exercise test – a procedure that must be done in a hospital, or similar clinical setting, under the auspices of a physician.

In addition to the problems associated with predicting maximum HR, heart rate is also affected by factors completely unrelated to exertion. Some of these factors include medication, the physical condition of the subject and test anxiety. Unless the effects of these factors are accounted for, the use of heart rate as an index is degraded even more as an index of exertion or cooperation during a lifting task. While the authors believe that estimates of maximum HR are appropriate for ensuring the safety of clients during functional testing, the use of this physiological response as an index of consistency of effort is a questionable concept.

4.8. *Impact of hopes and fears: the human element*

It would be completely disingenuous to state that any professional judgment made on the basis of a VEE cannot be impacted by the hopes and fears of the evaluator. For example, an evaluator might *hope* that the subject being tested will be cooperative throughout an evaluation because the evaluator has empathy for the subject. An evaluator could conceivably *hope* for failure because of a perverse financial incentive to maintain a commercial relationship with a referral source. An evaluator might *fear* that asking a subject to lift more weight might result in an injury – or *fear* being unjustifiably accused of hurting the subject if additional weight is added to a workload. Hopes and fears such as these are the human elements that may impact the results of a VEE "in the real world."

4.9. Weakness of a heuristic approach

It may well be argued that a heuristic approach that would include the ability for the subjects to observe the subjects "live" would provide additional information that would improve the model. In a past study involving subjects viewing muted videos of individuals engaged in lifting activities, the authors proposed that specifically mentioned the absence of *sound* as a factor that might affect the accuracy of a *visual* assessment of effort [18]. Observations relative to sounds, respiration, increased blood flow to the face, or even comments made by the person being tested *might* increase the level of accuracy of the estimations in some cases, but not necessarily in a manner that would be perceived by all evaluators. In some instances, such as might occur during a deliberately-deceptive performance, the variable of sound and other non-visual phenomena might simply cause additional error. Therefore, the question as to which impressions are correct and which are not correct would remain a salient and unresolved issue. Equally certain, there is no way by which conclusions from such a body of data can be "interpreted" in a standardized manner.

4.10. Similarity to past published studies

In a previous study, accuracy in the identification of maximum lifts was 46–53% when kinesiophysical operational definitions were applied to performances of asymptomatic subjects [30]. In that study, accuracy fell to 5–7% with regard to the identification of lifts performed at maximum levels of exertion when chronic back pain patients performed a lifting task. The findings presented in this study, therefore, are in substantial agreement with those findings with regard to identification of the correct identification of repetitive lifts at the 100% level.

In a study from the field of experimental psychology published nearly 30 years ago, the researchers reported findings that are also substantially similar to those reported in this study [31]. In the study, which investigated the topic of deception during a lifting task, subjects viewed videotaped performances of actors giving a good effort and feigning difficulty while lifting three different workloads (6.5 kg, 11.5 kg and 19.0 kg). Sensitivity to feigned difficulty was 33%, 77% and 83% for these three workloads, respectively. Specificity in identifying good effort was 83%, 80% and 77%, respectively. Notably, *the observers in this study were university students* who had no formal training in classifying effort during a lifting task. Although the task of the student observers was not identical to that of the subjects in the present study, the methodologies and findings are similar enough to conclude that the data presented in the present study are not startling, revolutionary findings.

4.11. Strengths and weaknesses of this study

The number of actors in this study, four, was small. But the range of ages and body types of the actors, as well as their medical histories, provide a reasonably good representation of the kinds of subjects who might conceivably appear for functional testing.

One possible weakness of the study is the data were collected independent of the control of the authors, with the exception of 10 subjects for whom data was entered in the presence the third author (RT) and seven subjects for whom data was gathered by another party (see Acknowledgments). It would be possible to argue that one cannot ensure the subjects did not rewind the videos or, in fact, that they did not even watch the videos. The striking similarity in the average number of correct estimations as well as the standard deviations for all groups for all activities is brought to the reader's attention. Such similarities would be unlikely in the event that the subjects did not, in fact, follow the written instructions.

This study has a large sample size in terms of the number of estimations made by the subjects. Ideally, additional therapists could have been recruited to take part in the data collection. In fact, one author of the present study (DS) sent approximately 150 email invitations to participate in the study to individuals listed on specific web sites that promote a network providing proprietary functional capacity evaluation services. Invitations were also sent to various officials of a prominent national organization of therapists, announcing the study and the opportunity for members of their organization to participate by accessing the test online. Unfortunately, none of the individuals receiving the invitations participated. Lacking any evidence that the results would have been different if additional subjects had been recruited for participation or if specific therapists would have participated – the authors contend the results reported in this study would be substantially similar if other individuals conducted the same experiment. Even if the results would be substantially different, the numerous weaknesses of VEE methodology would persist. Furthermore, this fact would remain: It is impossible to know with any reasonable degree of certainty which VEE estimate is correct – or incorrect – for any given estimation of effort or cooperation.

4.12. The expert witness culture and legal considerations

The high error rate in the classification of relative exertion during the repetitive lifts and effort during the incomplete lifts reported here is explainable in the context of the several and substantial weaknesses discussed herein. The failure to consider these weaknesses may very well be the origins of credibility problems associated with FCEs and discussed in trade publication editorials [24,35]. The failure to address the substantial problems in the visual assessment of effort may contribute substantially to the "expert witness" culture in this field and the litigious nature of compensable injury cases. One published article has discussed the issue of the expert witness culture [32]. In such an environment, questions relative to function and cooperation are often "settled" in a legal system in which competing interests are represented by opposing experts. The outcome is subject to a legal process which is conceivably affected greatly by the relative skills of the attorneys, and the judgment (good or bad) of arbitrators, judges and juries. When referring to the admissibility of testimony in medical-legal cases, one Supreme Court stated, "The subject of an expert's testimony must be 'scientific knowledge.' The adjective 'scientific' implies a grounding in the methods and procedures of science. Similarly, the word 'knowledge' connotes more than subjective belief or unsupported speculation. The term applies to any body of known facts or to any body of ideas inferred from such facts or accepted as truths on good grounds (fourth page)" [7]. The standard, if applied to testimony regarding a VEE, may very well result in a court's refusal to accept the testimony of the expert witness.

5. Conclusions

There are multiple conceptual, practical, methodological, physical and perceptual explanations for the inherent inaccuracy of the VEE. Education (including specialized training), years of experience in the field, and number of tests administered were not correlated with accuracy in estimating the relative level of exertion or the degree of cooperation in this study. Untrained lay subjects and untrained healthcare professionals were as accurate as the therapists who had training in the administration of lifting evaluations and had experience in conducting such tests. That is to say all groups produced equally unimpressive results with regard to accuracy – which marginally exceeded the level of chance. The authors question the validity of using the VEE in the assessment of individuals who are involved in medical-legal cases. Such methodology clearly does not meet the legal standard for admissibility as "scientific knowledge," as defined in Daubert v. Dow, 509 U.S. 579 (1993).

Acknowledgements

The authors would like to thank Catherine Wass and Stuart Hamilton (http://www.hamiltonsfitness.co. uk, Hamiltons Health and Fitness Ltd) for their kind permission to use Ms. Wass' photo in this article. The authors also wish to express their grateful appreciation to Sam Tsang and Dave Draper, http://www.davedraper.com (IronOnLine), for the use of Mr. Tsang's photograph for this article. Lastly, the authors wish to thank Theresa Delvo, PT, West Central Vice Chair for the Illinois Physical Therapy Association and practicing therapist at Orthopedic Center of Illinois, Springfield, Illinois for providing seven sets of data for this study.

References

[1] S. Anstis, Picturing peripheral acuity, *Perception* **27** (1998), 817–825.
[2] S. Brouwer, M.F. Reneman, P.U. Dijkstra, J.W. Groothoff, J.M. Schellekens and L.N. Göeken, Test-retest reliability of the Isernhagen Work Systems functional capacity evaluation in patients with chronic low back pain, *J Occup Rehabil* **13** (2003), 207–218.
[3] P.N. Brubaker, F.J. Fearon, S.M. Smith, R.J. McKibben, J. Alday, S.S. Andrews, E. Clarke and G.L. Shaw, Sensitivity and specificity of the Blankenship FCE system's indicators of submaximal effort, *J Orthop Sports Phys Ther* **37** (2007), 161–168.
[4] S.H. Ciriello, S.H. Snook, A.C. Blick and P.L. Wilkinson, The effects of task duration on psychophysically-determined maximum acceptable weights and forces, *Ergonomics* **2** (1990), 187–200.
[5] V.M. Ciriello, S.H. Snook and G.J. Hughes, Further studies of psychophysically determined maximum acceptable weights and forces, *Hum Factors* **35** (1993), 175–186.
[6] S. Coren, L.M. Ward and J.T. Enns, *Sensation and Perception,* 6e. Hoboken, NJ: John Wiley, 2004.
[7] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993).
[8] M.J. Durand, P. Loisel, S. Poitras, R. Mercier, S.R. Stock and J. Lemaire, The interrater reliability of a functional capacity evaluation: the physical work performance evaluation, *J Occup Rehabil* **14** (2004), 119–129.
[9] L. Feeler, *WorkSTEPS Class Manual*, 2008, Austin, TX.

[10] A. Garg, D.B. Chaffin and G.D. Herrin, Effects of lifting frequency and technique on physical fatigue with special reference to psychophysical methodology and metabolic rate, *Am Ind Hyg Assoc J* **40** (1979), 894–903.
[11] V. Gouttebarge, H. Wind, P.P. Kuijer, J.K. Sluiter and M.H. Frings-Dresen, Intra- and interrater reliability of the Ergo-Kit functional capacity evaluation method in adults without musculoskeletal complaints, *Arch Phys Med Rehabil* **86** (2005), 2354–2360.
[12] V. Gouttebarge, H. Wind, P.P. Kuijer, J.K. Sluiter and M.H. Frings-Dresen, Reliability and agreement of 5 Ergo-Kit functional capacity evaluation lifting tests in subjects with low back pain, *Arch Phys Med Rehabil* **87** (2006), 1365–1370.
[13] V. Gouttebarge, H. Wind, P.P. Kuijer and M.H. Frings-Dresen, Reliability and validity of Functional Capacity Evaluation methods: a systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen work system, *Int Arch Occup Environ Health Epub* **77** (2004), 527–537.
[14] D.P. Gross and M.C. Battié, Reliability of safe maximum lifting determinations of a functional capacity evaluation, *Phys Ther* **82** (2002), 364–371.
[15] P. Hallett, Eye movements, in: *Handbook of Perception and Human Performance: Volume 1 Sensory Processes and Perception,* K. Boff, L. Kaufman and J. Thomas, eds, Toronto: Wiley-Interscience, 1986, pp. 10-1 to 10-112.
[16] A.P. Hodselmans, P.U. Dijkstra, C. van der Schans and J.H. Geertzen, Test-retest reliability of psychophysical lift capacity in patients with non-specific chronic low back pain and healthy subjects, *J Rehabil Med* **39** (2007), 133–137.
[17] S. IJmker, E.H. Gerrits and M.F. Reneman, Upper lifting performance of healthy young adults in functional capacity evaluations: a comparison of two protocols, *J Occup Rehabil* **13** (2003), 297–305.
[18] S.L. Isernhagen, D.L. Hart and L.M. Matheson, Reliability of independent observer judgments of level of lift effort in a Kinesiophysical functional capacity evaluation, *Work* **12** (1999), 145–150.
[19] M.A. Jay, J.M. Lamb, R.L. Watson, I.A. Young, F.J. Fearon, J.M. Alday and A.G. Tindall, Sensitivity and specificity of the indicators of sincere effort of the EPIC lift capacity test on a previously injured population, *Spine* **25** (2000), 1405–1412.
[20] D.E. Lechner, J.R. Jackson, D.L. Roth and K.V. Straaton, Reliability and validity of a newly developed test of physical work performance, *J Occup Med* **36** (1994), 997–1004.
[21] M. Lemstra, W.P. Olszynski and W. Enright, The sensitivity and specificity of functional capacity evaluations in determining maximal effort: a randomized trial, *Spine* **29** (2004), 953–959.
[22] H. Lygren, T. Dragesund, J. Joensen, T. Ask and R. Moe-Nilssen, Test-rest reliability of the progressive isoinertial lifting evaluation (PILE), *Spine* **30** (2005), 1070–1074.
[23] L.N. Matheson, V. Mooney, J.E. Grant, M. Affleck, H. Hall, T, Melles, R.L. Lichter and G. McIntosh, *Spine*, A test to measure lift capacity of physically impaired adults. Part 1–development and reliability testing, **20** (1995), 2119–2129.
[24] J. Mecham, Under closer scrutiny: functional capacity evaluations must be evidence based and valid, *Advance for Directors in Rehabilitation*, June 2008, pp. 47–48.
[25] A. Mital, The psychophysical approach in manual lifting – a verification study, *Human Factors* **25** (1983), 485–491.
[26] M.F. Reneman, S.M. Jaegers, M. Westmaas and L.N. Göeken, The reliability of determining effort level of lifting and carrying in a functional capacity evaluation, *Work* **18** (2002), 23–27.
[27] M.F. Reneman, P.U. Dijkstra, M. Westmaas and L.N. Göeken, Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation, *J Occup Rehabil* **12** (2002), 269–275.
[28] M.R. Reneman, S. Brouwer, A. Meinema, P.U. Dijkstra, J.H. Geertzen and J.W. Groothoff, Test-retest reliability of the Isernhagen Work Systems functional capacity evaluation in healthy adults, *J Occup Rehabil* **14** (2004), 295–305.
[29] M.R. Reneman, Introduction to the special issue on functional capacity evaluations: from expert based to evidence based, *J Occup Rehabil* **13** (2003), 203–205.
[30] M.F. Reneman, A.S. Fokkens, P.U. Dijkstra, J.H. Geertzen and J.W. Groothoff, Testing lifting capacity: validity of determining effort level by means of observation, *Spine* **30** (2005), E40–E46.
[31] R.A. Robergs and R. Landwehr, The surprising history of the "HRmax = 220-age equation, *J Exercise Phys Online* **5** (2002). May, pp. 1–10. [Available from: http://www.faculty.css.edu/tboone2/asep/May2002JEPonline.html].
[32] S. Runeson and G. Frykholm, Kinematic specification of dynamics as an informational basis for person-and-action perception: expectation, gender, recognition and deceptive intent, *J Exper Psych* **112** (1983), 585–615.
[33] G. Rustenburg, P.P. Kuijer and M.H. Frings-Dresen, The concurrent validity of the ERGOS Work Simulator and the Ergo-Kit with respect to maximum lifting capacity, *J Occup Rehabil* **14** (2004), 107–118.
[34] D.J. Simons and C.F. Chabris, Gorillas in our midst: sustained inattentional blindness to dynamic events, *Perception* **28** (1999), 1059–1074.
[35] G. Simons, Credibility crisis in FCEs, *Physical Therapy Products*, October 2006, 3.
[36] S.H. Snook and C.H. Irvine, Maximum acceptable weight of lift, *Am Indus Hyg Assoc J* **28**(1967), 322–329.
[37] S.H. Snook, C.H. Irvine and S.F. Bass, Maximum weights and work loads acceptable to male industrial workers, *Am Ind Hyg Assoc J* **31** (1970), 579–586.
[38] S.H. Snook and V.M. Ciriello, Maximum weights and work loads acceptable to female workers, *J Occup Med* **8** (1974), 527–534.
[39] R.L. Smith, Therapists' ability to identify safe maximum lifting in low back pain patients during functional capacity evaluation, *J Orthop Sports Phys Ther* **19**(5) (1994), 277–281.
[40] R. Soer, E.H. Gerrits and M.F. Reneman, Test-retest reliability of a WRULD functional capacity evaluation in healthy adults, *Work* **26** (2006), 273–280.
[41] R. Soer, B.J. Poels, J.H. Geertzen and M.F. Reneman, A comparison of two lifting assessment approaches in patients with chronic low back pain, *J Occup Rehabil* **16** (2006), 639–646.
[42] R.J. Smeets, H.J. Hijdra, A.D. Kester, M.H. Hitters and J.A. Knottnerus, The usability of six physical performance tasks in a rehabilitation population with chronic low back pain, *Clin Rehabil* **20** (2006), 989–997.
[43] J.D. St. James, D.W. Schapmire, R. Townsend, L. Feeler and J. Kleinkort, Simultaneous, bilateral hand strength testing in a client population, part II: relationship to a distraction-based lifting evaluation, *Work* **37**(4) (1 Jan 2010), 395–403.
[44] N.L. Tuckwell, L. Straker and T.E. Barrett, Test-retest reliability on nine tasks of the Physical Work Performance Evaluation, *Work* **19** (2002), 243–253.