Quality of functional capacity evaluation tests: a clinimetric approach

Gouttebarge, V.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 6

Criterion-related validity of Functional Capacity Evaluation lifting tests on future work disability risk and return to work in the construction industry

Gouttebarge V, Kuijer PPFM, Wind H, van Duivenbooden C, Sluiter JK, Frings-Dresen MHW. Criterion-related validity of Functional Capacity Evaluation lifting tests on future work disability risk and return to work in construction industry Submitted
Abstract

Objectives
To assess the concurrent and predictive (criterion-related) validity of the five Ergo-Kit (EK) FCE lifting tests in construction workers on sick leave due to musculoskeletal disorders (MSDs).

Methods
Six weeks (baseline, t₀), six months (t₁) and one year (t₂) after the first sick leave day due to MSDs, construction workers underwent two isometric and three dynamic EK FCE lifting tests, and completed the Instrument for Disability Risk (IDR) for future work disability risk. Concurrent validity was assessed by the associations between the scores of the EK FCE lifting tests and the IDR outcomes, using Pearson Correlation coefficients (r) and associated proportions of variance (PV). Predictive validity was assessed by the associations between the EK FCE lifting test scores at baseline, and the IDR outcomes six months and one year later, using r, PV and area under receiver operating characteristic curve (AUC). In addition, the predictive validity of the EK FCE lifting tests on the total number of days on sick leave until full durable return to work (RTW) was evaluated with Cox regression analysis.

Results
Concurrent validity with future work disability risk was poor for the two isometric EK FCE lifting tests (-0.15 ≤ r ≤ 0.04) and moderate at t₁ and/or t₂ for the three dynamic EK FCE lifting tests (-0.47 ≤ r ≤ -0.31). Only the Carrying lifting strength test showed moderate and acceptable predictive validity on future work disability risk, especially at t₁ (r = -0.39; AUC = 0.72). Cox regression analyses revealed that two out of the five EK FCE lifting tests predicted durable RTW significantly, but only weakly.

Conclusions
Criterion-related validity with future work disability risk was poor for the two isometric EK lifting tests and moderate for the three dynamic lifting tests, especially the Carrying
lifting strength test. Predictive validity on durable RTW was significant although weak in two dynamic EK FCE lifting tests.
Chapter 6

Criterion-related validity of EK FCE lifting tests

Introduction

Edwin Smith’s Surgical Papyrus, roughly written in 1700 BC, is the world’s earliest known document that acknowledged signs of work-related musculoskeletal disorders (MSDs) in construction workers that arose from the imposing Egyptian pyramids construction projects (1). Nowadays, MSDs are a major burden on working populations, health systems and social care programs worldwide. In construction industries all over the world, MSDs are the primary reason for long-term sickness absence and related work disability, and the incidence of MSDs is strongly associated with manual material handling, especially lifting (2,3). In 2005, for the construction industry of the United States of America, overexertion when lifting caused 42% of the work-related MSDs with associated days away from work, while lifting was responsible for 21% of work compensation due to MSDs (4,5).

In order to reduce sick leave and work compensation costs due to MSDs, and to facilitate the return to work (RTW) process, occupational and insurance physicians need to assess the physical ability or inability to work (“physical work-ability”) of an injured worker, in particular, the ability to perform safe lifting among construction workers. In the Netherlands, physicians working either in RTW or disability claims do not possess many instruments to assess physical work-ability but they have a positive view on the utility of complementary information derived from the Functional Capacity Evaluation (FCE) (6). The FCE was designed to offer comprehensive performance-based assessments to measure the current physical work-ability of workers with or without MSDs (7-9). The Ergo-Kit (EK) is a FCE method that relies on a battery of standardized tests that assess work-related activities, such as standing, walking, lifting, carrying and reaching (10). As lifting ability is one of the most important components of heavy physical work, especially in the construction industry, the EK FCE lifting tests in particular could be seen as useful tools for the assessment of physical work-ability in the construction industry.

Validation of instruments is challenging and is the main topic of interest when it comes to the evaluation of the quality of an instrument’s measurements i.e., its clinimetric
properties. Without the assessment of validity, it cannot be claimed that what is purportedly being measured is what is truly being measured (7,11,12). Therefore, before one can administer the EK FCE lifting tests in occupational health care settings in the construction industry, the validity of the tests must be assessed. Among the different validity types, criterion-related validity is especially relevant for functional assessments (13-15). Criterion-related validity, subdivided into concurrent and predictive validity, describes how the evaluated test relates to another existing instrument measuring the same concept (or partially the same concept), ideally a gold standard showed to be reproducible and valid (13-15). Concurrent validity refers to the relation between the two instruments concurrently, meaning nearly at the same time, while predictive validity refers to the relation between two instruments, where the existing instrument is measured later on time (13-15). When no gold standard is available, as in the case of the assessment of physical work-ability (16,17), a well-grounded reference test (also referred as a silver standard) measuring an affiliated relevant concept and accepted in practice is commonly used as an alternative (18,19). In the Netherlands, the Instrument for Disability Risk (IDR) is an established and accepted instrument for identifying construction workers at risk for work disability over a two-year period (20,21). The IDR is a questionnaire assessing the status of four risk factors of future work disability in construction industry: age, sickness absence, musculoskeletal complaints and work ability (based on the Work Ability Index) (22). The IDR is appropriate as a reference test because it is a well-grounded instrument that is accepted and used in the construction industry and an instrument that measures future work disability risk, an affiliated concept of physical work-ability. Furthermore, the EK FCE tests were found to provide occupational professionals with complementary information that was useful when they made judgments of workers’ physical work-ability to aid the RTW process (23-25). Hence, the time until durable RTW (i.e., the number of days on sick leave until full durable RTW) seems another relevant affiliated concept that could be used in a validity study of the EK FCE lifting tests.

Thus, the three aims of the present study were to assess (a) the concurrent validity of the EK FCE lifting tests and future work disability risk in construction workers, (b) the predictive validity of the EK FCE lifting tests on future work disability risk in construction
workers on sick leave due to MSDs, and (c) the predictive validity of the EK lifting tests on time until durable RTW in construction workers on sick leave due to MSDs.

**Material and methods**

**Design**
A longitudinal within-subject design with a one-year follow-up period was conducted to evaluate the concurrent and predictive (criterion-related) validity of five EK FCE lifting tests.

**Participants and recruitment procedures**
From a nationwide list obtained from the largest occupational health and safety service in the Dutch construction sector, construction workers on sick leave for three to four weeks were contacted by phone by the first author. If a worker expressed interested to participate, detailed written information on the study procedure was sent and signed statements of informed consent were obtained. A sample size calculation was performed for our research questions ([a] 2-tailed t test with $\alpha = 0.05$ and power = 0.80; [b] confidence level of 0.95, correlation coefficient set at .50 and limit at .30), and it indicated that a minimum of 50 subjects were required at the end of our one-year follow-up period. To take drop-outs during follow-up into account, we strived to include 75 participants at baseline, based on the following inclusion criteria: (1) performing heavy physical work in the construction industry, (2) age between 18 and 55 years, and (3) on sick leave for the last 6 weeks (± 1 week) due to MSDs. Participants were free to withdraw from the study at any time.

**Ergo-Kit FCE lifting tests**
Five EK FCE lifting tests that were found to be reliable in adults without MSDs and with low back pain (26,27) were selected for this study. They included two isometric
lifiting tests, Back-torso lift test (Btlt) and Shoulder lift test (Slt), as well as three dynamic lifting tests: Carrying Lifting Strength Test (Clst), Lower Lifting Strength Test (Llst) and Upper Lifting Strength Test (Ulst). Table 1 presents the descriptions and outcomes of the five EK FCE lifting tests. Standardized procedures were performed as described in the EK handbook (10). The assessment of the five EK FCE lifting tests took approximately 30 minutes and was done by certified raters.

**Instrument for Disability Risk (IDR)**

In the present study, the IDR was selected as the reference test. In the Netherlands, the construction industry has developed this construction-industry specific instrument to identify workers at risk for work disability over a two-year period (20,21). Assessing four risk factors for work disability in the construction industry (i.e., age, work ability, sickness absence and musculoskeletal complaints), the IDR score is calculated from responses to nine questions (see Appendix). The IDR provides two types of outcomes: (1) a binomial outcome, having an increased risk for work disability or not, and (2) a risk of work disability (percentage). A percentage of 38 or more has been chosen in expert consensus meetings as the cut-off point for an increased risk of work disability over 2 years (20,21).
Return to work (RTW)
In the present study, time to durable RTW was defined as the duration of work absenteeism due to MSDs in calendar days from the first day on sick leave until the first day of returning fully to the worker’s own work or other work for a period of at least four weeks (28). RTW was registered throughout the one-year follow-up period by the occupational health and safety service in the construction sector.

Study procedures
Six weeks (baseline, $t_0$), six months ($t_1$) and one year ($t_2$) after the first sick leave day, subjects were assessed on five EK FCE lifting tests and were asked to complete the IDR, during the occupational physician consultation at $t_0$, at $t_1$ and $t_2$ at home. To guarantee that the time interval between the two assessments (i.e., the EK FCE lifting tests and IDR) was as short as possible, participants who did not return the IDR questionnaire within three days after their assessment on the EK FCE lifting tests were again contacted by phone. This study was performed in accordance with the Helsinki Declaration (1964) and received approval from the Medical Ethics Committee of the Academic Medical Center in Amsterdam, the Netherlands.

Data analyses
Only the participants included at baseline and who completed the three assessments without any missing value(s) during the one-year follow-up period were included in the analyses. All data analyses were performed using the statistical analysis software SPSS 14.0 for Windows. At $t_0$ (baseline), $t_1$ and $t_2$, descriptive statistics (means, standard deviations and ranges) were calculated for each of the EK FCE lifting tests and the IDR outcome.

Concurrent validity was determined by assessing the relationship at $t_0$ (baseline), $t_1$ and $t_2$ between the five EK FCE lifting tests scores and the IDR outcomes. Predictive validity of the EK FCE lifting tests on future work disability risk was evaluated by assessing the associations between the five EK FCE lifting tests scores at $t_0$ (baseline)
and the IDR outcomes at both $t_1$ and $t_2$. For both concurrent and predictive validity on future work disability risk, Pearson’s correlation coefficients ($r$) were calculated and the associated proportions of variance (PV) were calculated ($PV = 100 \times r^2$) to illustrate how much variance in one variable (EK FCE lifting test) may be explained by the variance of the other variable (IDR) (29). For concurrent and predictive validity, correlations greater than .50 are considered as good, between .30 and .50 as moderate, and lower than .30 as poor (29).

Furthermore, the ability of the EK FCE lifting tests at $t_0$ (baseline) to predict the outcomes of the IDR at $t_1$ and at $t_2$ was measured using the area under the receiver operating characteristic curve (AUC) (15,30). Therefore, we used the cut-off point of 38% set by Dutch experts for a high risk for work disability (20,21). The AUC can be interpreted as follows: $0.7 \leq \text{AUC} \leq 0.8$ as acceptable, $0.8 < \text{AUC} \leq 0.9$ as excellent, and $\text{AUC} > 0.9$ as outstanding (31).

Predictive validity of the EK FCE lifting tests at $t_0$ (baseline) on time to durable RTW (number of days on sick leave) during the one-year follow-up was evaluated by conducting Cox proportional hazards regression analysis (30). For this analysis, the number of days on sick leave until durable return to work was calculated from six weeks after the first sick leave day, which was the time of the baseline assessment with the EK FCE lifting tests. Cox regression analysis was performed to identify whether the EK FCE lifting tests (independent variables or covariates) were separate predictive factors for time to durable RTW (dependent variable). Beta ($\beta$), hazard ratio (HR), 95% confidence interval (CI) and p-values were calculated and presented (30). Interpretation was based on the level of significance (p-values) and on the HR, that can be interpreted in a similar manner to odds ratio: values above one indicate a raised hazard for durable RTW, values below one indicate a decreased hazard and values equal to one indicate there is no increased or decreased hazard (32).
Results

Participant’s characteristics
Sixty construction workers completed the three assessments without any missing information during the one-year follow-up period. At baseline, all 60 participants were on sick leave due to MSDs, with the upper extremity MSDs accounting for 17% of the main diagnoses, the lower extremity for 28%, the back for 30%, and a combination of MSDs for the remaining 25%.

Participants were assessed on the EK FCE lifting tests in 15 different locations in the Netherlands, depending on their home addresses. Among the participants, carpentry was most frequent occupation (37%). From the 60 participants included at baseline in the study and on sick leave (+/- six weeks), 47 returned to work six months later (t1;78%) and 51 returned one year later (t2;85%). Nine participants were still on sick leave after the one-year follow-up period.

The baseline characteristics of the 60 participants are presented in Table 2. The participants’ mean age was 42 years, their mean height was 182 centimetres and their mean bodyweight was 86 kilograms, and their average number of days on sick leave (duration of sickness absence) was 146 days, ranging from 42 days to one year (at which point participants were censored).

Concurrent validity
Table 2 presents the outcomes at t0, t1, and t2 of the five EK FCE lifting tests and the IDR. The associations between the five EK FCE lifting tests scores and the IDR outcomes are presented in Table 3. At baseline weak associations were found between scores of the isometric and dynamic EK FCE lifting tests and the IDR outcomes (-0.17 ≤ r ≤ 0.07). At t1 and t2, the associations between the scores of the two isometric EK FCE lifting tests and the IDR outcomes were also weak. Moderate associations (p < .01) at t1 and/or t2 were found between the outcomes of the three dynamic EK FCE lifting tests and the IDR, with an upper value of r = -0.47 (p < .01) for the association at
Table 2: Means, standard deviations and ranges of age, height, bodyweight and outcomes of the Ergo-Kit FCE lifting tests and Instrument for Disability Risk (IDR) at t0 (baseline), t1 and t2 (N = 60).

<table>
<thead>
<tr>
<th></th>
<th>Baseline t0</th>
<th>t1</th>
<th>t2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Range</td>
</tr>
<tr>
<td>Age (years)</td>
<td>42</td>
<td>9</td>
<td>18 - 55</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>182</td>
<td>8</td>
<td>168 - 198</td>
</tr>
<tr>
<td>Bodyweight (kg)</td>
<td>86.4</td>
<td>32.2</td>
<td>21.0 - 152.5</td>
</tr>
<tr>
<td>Back-torso lift test (kg)</td>
<td>43.9</td>
<td>18.5</td>
<td>2.5 - 88.5</td>
</tr>
<tr>
<td>Shoulder lift test (kg)</td>
<td>36.2</td>
<td>13.3</td>
<td>10.0 - 75.0</td>
</tr>
<tr>
<td>Carrying lifting strength test (kg)</td>
<td>22.1</td>
<td>8.6</td>
<td>5.0 - 50.0</td>
</tr>
<tr>
<td>Lower lifting strength test (kg)</td>
<td>42.7</td>
<td>16.6</td>
<td>9.0 - 65.0</td>
</tr>
</tbody>
</table>

N, number of subjects; SD, standard deviation; cm, centimeter; kg, kilogram;

Table 3: Correlations (r) and proportions of variance (PV) between the outcomes of the Instrument for Disability Risk (IDR) and the Ergo-Kit FCE lifting tests at t0 (baseline), t1 and t2 (N = 60).

<table>
<thead>
<tr>
<th>Variables</th>
<th>t0 (r)</th>
<th>PV (r)</th>
<th>t1 (r)</th>
<th>PV (r)</th>
<th>t2 (r)</th>
<th>PV (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-torso lift test</td>
<td>-0.15</td>
<td>2.25</td>
<td>-0.15</td>
<td>2.25</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Shoulder lift test</td>
<td>-0.07</td>
<td>0.49</td>
<td>-0.05</td>
<td>0.25</td>
<td>0.04</td>
<td>0.16</td>
</tr>
<tr>
<td>Carrying lifting strength test</td>
<td>-0.17</td>
<td>2.89</td>
<td>-0.47**</td>
<td>22.09</td>
<td>-0.33**</td>
<td>10.89</td>
</tr>
<tr>
<td>Lower lifting strength test</td>
<td>-0.17</td>
<td>2.89</td>
<td>-0.36**</td>
<td>12.96</td>
<td>-0.31*</td>
<td>9.61</td>
</tr>
<tr>
<td>Upper lifting strength test</td>
<td>-0.12</td>
<td>1.44</td>
<td>-0.42**</td>
<td>17.64</td>
<td>-0.23</td>
<td>5.29</td>
</tr>
</tbody>
</table>

N, number of subjects; *, p < .05; **, p < .01
Table 4: Predictive validity of the Ergo-Kit FCE lifting tests at t₀ (baseline) on the Instrument for Disability Risk (IDR) at t₁ and t₂: correlations (r), proportions of variance (PV) and Area under the curve (AUC) (N = 60).

<table>
<thead>
<tr>
<th>Variables</th>
<th>IDR at t₀ r</th>
<th>PV</th>
<th>AUC</th>
<th>IDR at t₁ r</th>
<th>PV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-torso lift test at t₀</td>
<td>-0.14</td>
<td>1.96</td>
<td>0.50</td>
<td>-0.10</td>
<td>1.00</td>
<td>0.53</td>
</tr>
<tr>
<td>Shoulder lift test at t₀</td>
<td>-0.04</td>
<td>0.16</td>
<td>0.45</td>
<td>-0.09</td>
<td>0.81</td>
<td>0.52</td>
</tr>
<tr>
<td>Carrying lifting strength test at t₀</td>
<td>-0.39**</td>
<td>15.21</td>
<td>0.72</td>
<td>-0.32*</td>
<td>10.24</td>
<td>0.68</td>
</tr>
<tr>
<td>Lower lifting strength test at t₀</td>
<td>-0.29*</td>
<td>8.41</td>
<td>0.67</td>
<td>-0.19</td>
<td>3.61</td>
<td>0.60</td>
</tr>
<tr>
<td>Upper lifting strength test at t₀</td>
<td>-0.19</td>
<td>3.61</td>
<td>0.62</td>
<td>-0.22</td>
<td>4.84</td>
<td>0.58</td>
</tr>
</tbody>
</table>

N, number of subjects; *, p < .05; **, p < .01

Table 5: Predictive validity of the Ergo-Kit FCE lifting tests on RTW (number of days on sick leave until return to work): Cox proportional hazards regression analysis (N = 60).

<table>
<thead>
<tr>
<th>Variables</th>
<th>β</th>
<th>Hazard ratio</th>
<th>95% CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-torso lift test</td>
<td>0.008</td>
<td>1.00</td>
<td>1.00 – 1.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Shoulder lift test</td>
<td>0.009</td>
<td>1.00</td>
<td>1.00 – 1.02</td>
<td>0.22</td>
</tr>
<tr>
<td>Carrying lifting strength test</td>
<td>0.030</td>
<td>1.03</td>
<td>1.00 – 1.05</td>
<td>0.01</td>
</tr>
<tr>
<td>Lower lifting strength test</td>
<td>0.045</td>
<td>1.05</td>
<td>1.02 – 1.07</td>
<td>0.00</td>
</tr>
<tr>
<td>Upper lifting strength test</td>
<td>0.027</td>
<td>1.03</td>
<td>1.00 – 1.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>

N, number of subjects; CI, confidence interval; P, p-value
t₁ between the Carrying lifting strength test and the IDR. Furthermore, proportions of variance at t₁ ranged from 13 to 22% for the three dynamic EK FCE lifting tests.

Predictive validity IDR
The relationships between the five EK FCE lifting tests scores at t₀ (baseline) and the IDR outcomes at t₁ and at t₂ are presented in Table 4. One dynamic EK FCE lifting test, the Carrying lifting strength test, had a moderate correlation with the IDR (-0.39 at t₁ and -0.32 at t₂), showing a moderate predictive validity on future work disability risk. In addition, an acceptable predictive ability of the Carrying lifting strength test for IDR outcomes was confirmed by an AUC value of 0.72 at t₁. Weak associations were found between the scores on the other four out of the five EK FCE lifting tests and the IDR outcomes, ranging from -0.04 to -0.29. In addition, the AUC of these four EK FCE lifting tests ranged between 0.45 and 0.67, showing poor abilities to predict the IDR outcomes at t₁ and t₂ from the EK FCE lifting test scores.

Predictive validity durable RTW
Table 5 shows the results of the Cox proportional hazards regression analyses, revealing that two out of the EK FCE lifting tests (Carrying and Lower lifting strength test) were significant (p ≤ 0.03) although weak (HR = 1.03; HR = 1.05) predictors of the number of days on sick leave until durable RTW. The predictive validity of the other three lifting tests on durable RTW was poor (p > 0.05).

Discussion
Using a longitudinal within-subject design, the aim of this study was to evaluate the criterion-related (concurrent and predictive) validity of five EK FCE lifting tests (two isometric and three dynamic) in construction workers who were on sick leave at baseline because of MSDs. Concurrent validity between the two isometric EK FCE
lifting tests and the IDR, the reference test for future work disability risk, was found to be poor while concurrent validity between the three dynamic EK FCE lifting tests and the IDR was moderate. One dynamic EK FCE lifting test, the Carrying lifting strength test, showed a moderate level of predictive validity on the IDR. The predictive validity of the other four out of the five EK FCE lifting tests on the IDR was poor. Furthermore, the predictive validity of the five EK FCE lifting tests on durable RTW (i.e., number of days on sick leave until full durable RTW) could not be established. Overall, the criterion-related validity with future work disability risk was poor for the two isometric EK lifting tests and moderate for the three dynamic lifting tests, especially the Carrying lifting strength test. The predictive validity on durable RTW was significant but poor in two out of the five EK FCE tests.

Conducting a validity study of an instrument inevitably entails some methodological and procedural considerations, especially with regard to the study population, the reference test, the design and the statistical analyses. First, the study population chosen in any validity study is essential in order to validate correctly the evaluated instrument or test. As FCEs strive to report physical work-ability, the selection of construction workers in our validity study seems relevant as construction workers perform jobs particularly exposed to manual material handling, which is strongly related to the occurrence of MSDs and to sick leave (2,3). In addition, among all manual material handling activities performed in the jobs of the construction industry, lifting is definitely a dominant activity (2,3), making our choice of construction workers in the assessment of the EK FCE lifting tests validity even more appropriate. In order to keep 60 participants in the study after the one-year follow-up period, 72 were included at baseline. From these 72, eight dropped out (11%) for the following assessment six months later, an additional one dropped out (1%) between the second and third assessment, and three participants (4%) had missing value(s) on the EK FCE tests during the one-year follow-up period. The main reason for drop out was that participants did not find any time or motivation to be assessed again on the EK FCE lifting tests because they already returned to work, or they suffered from a MSD that did not allow them to be assessed again with the EK FCE lifting tests according to our study timetable. Compared with the participants who remained in our longitudinal study, the 12 participants who had missing value(s) and
were therefore excluded in the analyses were slightly younger (mean age of 37 years old) and stayed longer on sick leave (169 days). Finally, 60 construction workers completed the three assessments without any missing value(s) during the one-year follow-up period (83%), which appears as a relatively high rate. One reason for participants remaining in the study could be the financial reward they received: in addition to the traveling expenses, they received 50 Euro per assessment (150 Euro for the whole study period) and were entered into a lottery for a traveler's cheque with a value of 1000 Euro. All in all, the use of construction workers and the few drop-outs are strengths of the present study, as it seemed legitimate to select such a population in the validation process of the EK FCE lifting tests.

Second, with regard to the reference test selected, we could put forth reasons to justify the selection of the IDR for our criterion-related validity study. The concept that is measured by the EK FCE lifting tests is physical work-ability. As no gold standard is available for physical work-ability (16,17), a well-grounded instrument, accepted and used in practice, measuring physical work-ability or an affiliated relevant concept had to be selected. Considering the use of construction workers as participants in our study, especially in the context of the Dutch construction industry, and the need to have a test that was affiliated with the concept of physical work-ability, our search for a reference test resulted in the IDR. The IDR is intended to be used in the case of construction workers to assess future work disability risk due to MSDs, which seems an acceptable affiliated concept for (physical) work-ability. Furthermore, within the nine questions of the IDR, physical work-ability is specifically addressed. It also indirectly assesses the respondent's lifting ability, as this activity is one of the most important for jobs in the construction industry. Thus, as no gold standard is available for physical work-ability, the IDR appears as a rational reference test to assess the criterion-related validity of the EK FCE lifting tests in the construction industry.

Third, to explore and establish relationships particularly between the outcome(s) of evaluated instrument(s) (i.e., independent variable(s) at baseline) and the outcome(s) of interest (i.e., dependent variable(s)) during a follow-up period, an observational prospective longitudinal study design was the best suited research design, even if
observational studies provide weaker empirical evidence than do experimental studies (33). In the present study, a longitudinal within-subject design was used to assess criterion-related validity, which seemed relevant for predictive validity. In addition, as stated earlier, loss to follow-up has not been a critical issue as the rate of dropouts was low in our one-year follow-up period. Furthermore, a strength of the design chosen for this study was the possibility to assess concurrent validity between the EK FCE lifting tests and the IDR at three different moments within one year, allowing a comparison over time of the concurrent validity and the evaluation of the durability of validity in a ‘changing’ population, that is, workers recovering from MSDs and sickness absence. In the present study, the concurrent validity, particular of the dynamic EK FCE lifting tests, with future work disability risk changed substantially between baseline and either the second or third assessments.

Finally, with regard to the analyses, different statistical tools have been applied. Pearson correlation coefficients were used and proportions of variance were calculated to assess the association between the five EK FCE lifting tests and the IDR. Furthermore, based on the cut-off point of 38% set by experts as a higher risk for work disability (20,21), the IDR binominal outcomes allowed us to also use the AUC in order to evaluated the ability of the EK FCE lifting tests to predict the IDR outcomes. For predictive validity, regression analyses are often used. In our study, the purpose was not to find a model or a combination of EK FCE lifting tests that would predict the IDR, but to examine whether each FCE test could predict the IDR. For that reason, no regression analyses were performed and only correlation coefficients were calculated. In addition, the calculation of the proportion of variance from the correlation coefficients has some similarities with the R-squares available through regression analyses. Cox proportional hazards regression analysis was also conducted for each EK FCE lifting test to analyse the independent predictive powers on the number of days on sick leave until full durable RTW. All in all, the statistical tools that have been used were appropriate for the research questions in this criterion-related validity study.

As FCEs has been recently a topic of interest (34), our results can be compared with other criterion-related validity studies. As in the present study, some authors tried to
assess the concurrent validity of FCE tests with self-reported questionnaires measuring disability-related concepts. Similar to our results, Reneman et al. found low to moderate levels of concurrent validity between the Isernhagen Work Systems (IWS) FCE lifting and carrying tests, and three self-reported disability questionnaires (Roland-Morris Disability questionnaire, Oswestry Back Pain Disability Scale and Quebec Back Pain Disability Scale), while Gross and Battie found fair-to-moderate associations between the IWS FCE carrying and lifting tests, and the Pain Disability Index and pain visual analogue scale (35-37). From this perspective, it can be suggested that physical work-ability, measured through the IWS or EK FCE, and self-reported questionnaires measuring disability-related concepts, can be seen as affiliated or related to each other. However, a study comparing concurrently the IWS and EK FCE lifting tests showed that both the FCEs produced different results, meaning that the IWS and EK can not be used interchangeably (38). In our study, the predictive validity of four out of the five EK FCE lifting tests on future work disability risk was not established and the durable RTW could not be predicted by the EK FCE lifting tests ($1.00 \leq \text{HR} \leq 1.03$). These results are in line with others studies (39-41) that dealt with the predictive value of FCEs on RTW, although these studies were conducted in a different context (i.e., work disability claim context). Gross and Battie found that a better lifting ability was weakly related to either faster or safer RTW. Using data from a prospective cohort study of subjects with chronic low back pain, the same authors also found that better FCE performance was slightly associated with RTW (39-41). An explanation for our results may be the quick evolution over time of the nature of the participants’ MSDs. Another possibility may be that the expectation that FCEs, which measure current physical work-ability, have prognostic value on future work-related concepts could be just too ambitious and not realistic.

From the results of this study, it seems that only one test out of the five EK FCE lifting tests evaluated, the Carrying lifting strength test, gives information that is moderately valid for the construction industry. The isometric EK FCE lifting tests, reflecting maximal lift capacity, did not give valid measurements with regard to future work disability risk and RTW. Lasting only a few seconds, the two isometric tests appeared less relevant for the work demands in the construction industry and this may partially
explain the results of our study. Compared with the other two dynamic EK FCE lifting tests, the Carrying lifting strength test reflects the largest number of activities such as gripping, lifting, bending, carrying and walking. Walking is especially responsible for the longer time needed for the assessment and seems relevant to the physical work demands of construction workers, which could be an explanation for its moderate association with future work disability risk. However, as it presents only a moderate evidence of criterion-related validity, the Carrying lifting strength test cannot be used solely for jobs exposed to manual material handling as a test used by occupational professionals working in health and safety services. In addition, the construct validity of the five EK FCE lifting tests, including this Carrying lifting strength test, was not supported (42).

Thus, it seems necessary to first evaluate whether the information from the Carrying lifting strength test, in combination with information provided by anamnesis, clinical examination and self-reported questionnaires, could have an added value for the judgment and decision making process of occupational professionals in their assessment of physical work-ability. If so, and only if so, the assessment of the Carrying lifting strength test could provide occupational professionals with useful and valid information on several activities in a rapid and efficient way, and it would also enhance the practicality of using FCEs to some extent. Indeed, FCEs practicality is known to be limited as they are often generic and time consuming. This practicality aspect has been logically a topic of interest for some authors as efforts have been made to increase the FCE practicality by selecting functional tests from the full FCE that may be appropriate for specific defined jobs (43-45). However, further research on shorter and more specific FCEs are still needed to support their application in occupational medicine for heavy physical jobs such as construction workers, firefighters or garbage collectors. Furthermore, gathering information from different sources such as self-reported questionnaires, clinical examination and performance-based testing (i.e., FCEs), could lead to an optimal assessment of current physical work-ability, and should be subject to further research.
Conclusion

Criterion-related validity with future work disability risk in sick-listed construction workers with MSDs was poor for the two isometric EK lifting tests and moderate for the three dynamic lifting tests, with the highest value for the Carrying lifting strength test. Good predictive validity of the five EK FCE lifting tests on durable RTW was not observed; although in two out of the five EK FCE tests it was significant, the validity was weak.

Acknowledgments

The authors would like to thank Marco van de Velde (‘ArboDuo’) for his support in this study. We are also grateful to all certified raters who assessed the EK FCE lifting tests and to all participants.

Funding

This study was financially supported by ‘Arbouw’ and ‘Stichting Instituut GAK’.
Appendix: Instrument for Disability Risk (IDR) (20,21).

1. How would you rate your current work ability compared with the lifetime best, where 0 is ‘not able to work’ and 10 is ‘best work ability ever’? (0-10 scale)

2. How would you rate your current work ability with regard to the physical work demands of your job?’ (5-points Likert scale)

3 How would you rate your current work ability with regard to the psychological work demands of your job? (5-points Likert scale)

4. From the following list of 51 diseases, give the number of current diseases you have that were diagnosed by a physician and / or diagnosed by your self. (number of diseases)

5. Give your estimation of work impairment due to diseases. (1-6 scale)

6. How many days were you on sick leave during the past year? (1-5 scale)

7. From your own judgment, do you think you will be working in your own job in two years? (3-points Likert scale)

8a. Lately, do you enjoy your daily life? (5-points Likert scale)
8b. Lately, have you been active and fit? (5-points Likert scale)
8c. Lately, have you had trust in the future? (5-points Likert scale)

9a. Do you have regular neck stiffness or pain? (binominal)
9b. Do you have regular stiffness or pain in the upper extremity? (binominal)
9c. Do you have regular back stiffness or pain? (binominal)
9d. Do you have regular stiffness or pain in the lower extremity? (binominal)
References

8. Strong S. Functional capacity evaluation: the good, the bad and the ugly. OT Now. 2002;Jan/Feb:5-9
23. Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. The complementary value of Functional Capacity Evaluation for physicians in assessing the physical work-ability of workers with musculoskeletal disorders. Submitted