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 7

General discussion
Background

As clinimetric properties must be evaluated before beginning to implement any test, the main objective of this thesis was to gather scientific evidence on the clinimetric properties, i.e. measurement quality, of Functional Capacity Evaluation (FCE) tests that are available and used in the Netherlands, focusing on whether FCE tests give reproducible outcomes in repeated measurements and whether the outcomes of FCE tests are valid for the assessment of physical work-ability in a return to work context. In this chapter, the main findings will be reported in the light of the six different research questions, and the selection of FCE tests will be covered. Then, the clinimetrics model illustrated in the general introduction (Chapter 1) will be discussed and methodological considerations given. A general conclusion will then be drawn, and implications for the use of FCE tests will also be suggested in addition to areas for future research.

Main findings

In the light of the six research questions formulated in this thesis (Chapter 1), the main findings flowing from the body of empirical studies are briefly reported in the following section.

Reproducibility of EK FCE tests

1. What is known in the international literature about the reliability of four FCE methods available in the Netherlands?

A systematic literature search performed using five electronic databases allowed us to identify studies on the reproducibility of several Isernhagen Work System tests, but we did not find any evidence of the reproducibility of the other three FCE methods (Blankenship System, Ergos Work Simulator and Ergo-Kit), demonstrating the need for studies of this clinimetric property (Chapter 2).
2. How reproducible are EK FCE tests in subjects without musculoskeletal complaints?

A within-subjects design explored the intra- and interrater reliability of seven EK FCE tests (two isometric EK FCE lifting tests: the Back-torso lift test and the Shoulder lift test; three dynamic EK FCE lifting tests: the Carrying lifting strength test, the Lower lifting strength test and the Upper lifting strength test; two EK FCE manipulation tests: the Forward manipulation test and the Lower manipulation test crouching) in adults without musculoskeletal complaints (MSCs). Intrarater reliability was high for the two isometric EK FCE lifting tests (intraclass correlation coefficient ICC ≥ 0.85), moderate for the three dynamic EK FCE lifting tests (0.69 ≤ ICC ≤ 0.75), and low for the two manipulation EK FCE tests (ICC ≤ 0.46). Interrater reliability was moderate (8-day time interval; 0.67 ≤ ICC ≤ 0.90) to high (4-day time interval; ICC ≥ 0.85) for the two isometric and the three dynamic EK FCE lifting tests, and low (8-day time interval; ICC ≤ 0.01) to moderate (4-day time interval; 0.74 ≤ ICC ≤ 0.78) for both manipulation EK FCE tests (Chapter 3).

3. How reproducible are EK FCE tests in subjects suffering from musculoskeletal complaints?

A within-subjects design was conducted on subjects suffering from low back pain (LBP) to explore the reproducibility, i.e. reliability and agreement between two raters, of five EK FCE lifting tests (two isometric EK FCE lifting tests: the Back-torso lift test and Shoulder lift test; three dynamic EK FCE lifting tests: the Carrying lifting strength test, the Lower lifting strength test and the Upper lifting strength test) that were found reliable in adults without MSCs. With a 3-day time interval between test and retest, the five EK FCE lifting tests were found to be highly reliable (0.84 ≤ ICC ≤ 0.97) in subjects suffering from LBP, while the level of agreement between the two raters was considered to be good. Standard errors of measurement (SEM) ranged from 5.0 to 8.6 kilograms (means from 37.6 to 65.9 kg; standard deviations from 18.3 to 39.5 kg) for the isometric EK FCE lifting tests and from 1.9 to 3.7 kilograms (means from 17.0 to 24.5 kg; standard deviations from 6.3 to 11.2 kg) for the dynamic EK FCE lifting tests (Chapter 4).
Validity of EK FCE tests

4. What is known in the international literature about the validity of four FCE methods available in the Netherlands?

Through a systematic literature search performed using five electronic databases, seven original articles on the validity of FCEs available in the Netherlands (Blankenship System, Ergos Work Simulator, Ergo-Kit and Isernhagen Work System) were identified, with two of these seven assessing the concurrent validity of the EK FCE lifting tests with the Ergos Work Simulator and Isernhagen Work System lifting tests. Our systematic literature review showed low to moderate levels of validity of the EK FCE tests, showing that this clinimetric property needs to be further studied (Chapter 2).

5. What is the construct validity of EK FCE tests in employees on sick leave due to MSDs?

The discriminative and convergent validity of five EK FCE lifting tests (two isometric EK FCE lifting tests: the Back-torso lift test and the Shoulder lift test; three dynamic EK FCE lifting tests: the Carrying lifting strength test, the Lower lifting strength test and the Upper lifting strength test) were assessed in construction workers on sick leave for musculoskeletal disorders (MSDs). Discriminative validity was assessed by using the Known Groups Method and forming two groups of construction workers (high risk and low risk for future work disability based on the Instrument for Disability Risk) that were expected to be different from each other. The expected differences between the two groups of construction workers were observed for four out of the five EK FCE lifting tests, although none of the five EK FCE lifting tests could significantly discriminate between the two groups (0.07 ≤ p ≤ 0.79). Convergent validity of the five EK FCE lifting tests with self-reported pain intensity and disability was found to be poor (-0.29 ≤ Pearson correlation coefficient r ≤ 0.05). In this study, construct validity of the five EK FCE lifting tests in construction workers on sick leave due to MSDs could not be supported (Chapter 5).
6. What is the criterion-related validity of EK FCE tests in employees on sick leave due to MSDs?

The concurrent and predictive validity of five EK FCE lifting tests (two isometric EK FCE lifting tests: the Back-torso lift test and the Shoulder lift test; three dynamic EK FCE lifting tests: the Carrying lifting strength test, the Lower lifting strength test and the Upper lifting strength test) with the Instrument for Disability Risk (IDR), our reference test for future work disability risk, was assessed in construction workers on sick leave due to musculoskeletal disorders (MSDs). Concurrent validity with future work disability risk at three time points during a one year period was found to be poor for the two isometric EK FCE lifting tests (-0.15 ≤ r ≤ 0.04) and moderate for the three dynamic EK FCE lifting tests (-0.47 ≤ r ≤ -0.31). Only one dynamic EK FCE lifting test, the Carrying lifting strength test, showed a moderate level of predictive validity on future work disability risk, especially six months after baseline (r =-0.39; Area under the curve AUC=0.72). Furthermore, the predictive validity of the five EK FCE lifting tests on the number of days on sick leave until full durable (for at least four weeks) return to work was not established (1.00 ≤ Hazard ratio HR ≤ 1.05); it was significant for two EK FCE lifting tests (Carrying and Lower lifting strength test), but was very weak. Therefore, in this study, a poor level of criterion-related validity was found for the two isometric EK lifting tests, while the three dynamic lifting tests showed some moderate evidence of criterion-related validity with future work disability risk, especially the Carrying lifting strength test (Chapter 6).

Selection of FCE tests

From our systematic literature review, several studies were found on the reproducibility and validity of the Isernhagen Work System (IWS) tests, especially the lifting and carrying tests. Intra- and interrater reliability was rated from moderate to high (ICC ≥ 0.75; agreement ≥ 87%; kappa k ≥ 0.68), while evidence of predictive validity on return to work was found by Matheson et al. (2002) but not by Gross et al. (2004) (Chapter 2). As no reliability study was found for the Blankenship System (BS), the Ergos Work
Simulator (EWS) or the Ergo-Kit (EK) and only limited evidence of validity was found for the Ergos Work Simulator (EWS) and the Ergo-Kit (EK), the results of our systematic literature search emphasized the need for empirical research on clinimetrics of these three FCEs (Chapter 2). For our first empirical study on reproducibility, we selected and used tests from the EK FCE rather than tests from the BS and EWS because of its practicality and availability in the Netherlands.

Practicality is a critical issue when it comes to the selection of functional tests; some authors even note that an FCE’s practicality is more important than statistical evidence of reliability and validity (1-3). The practicality of a test deals with aspects such as standardization of equipment and instructions, ease of administration, and complexity of scoring, reporting and interpreting the data, and is also related to its direct and indirect costs (4-6). As the EK FCE is strictly standardized and easily mobile, its practicality seems acceptable. Furthermore, while the BS and EWS were at the time of our first study only available in a few places in the Netherlands, the EK FCE is forthcoming in more than thirty locations in the Netherlands. This is also an important argument when it comes to the eventual implementation of the EK FCE because its wide availability in the Netherlands enables occupational professionals to refer patients nationwide to the EK FCE. Thus, based on practicality and availability, tests from the EK FCE were chosen for use in our empirical studies.

However, as the whole EK test protocol, based on 55 tests, is completed within approximately three hours, and as the same activity can be reflected through several tests, the EK FCE remains time-consuming and generic, giving redundant information for some activities: in other words, its practicality could be increased. In recent years, the practicality of FCEs has been a topic of interest, and some studies have focused on increasing FCE practicality by selecting functional tests from a full FCE method with regard to a defined job (3,7). In this approach, the required job demands are converted into a specific FCE protocol and compared to the functional physical abilities of a worker. In this job matching approach, the results of the selected FCE tests are used to establish whether a worker is capable of performing the required job demands in order to facilitate and empower the return to work process. In addition, such a selection from
a full FCE method enhances FCE practicality, but does not threaten the validity of the selected tests (8).

Analogously to the selection of FCE tests for a defined job, as several of the ‘physical agility’ EK FCE tests could be acknowledged as tests measuring the same activity, EK FCE tests were selected at the start of our first empirical study (Chapter 3) in order to enhance the practicality and to more efficiently assess the specific physical work-ability in patients suffering from a particular MSC. Based on literature and expert opinions, a three-step procedure was considered. First, three groups of MSCs, i.e. MSDs, were created: upper extremity MSC, back MSC and lower extremity MSC. Second, if a patient suffers from a particular MSC, several physical activities can be chosen that he or she might find difficult to perform. Activities that may be troublesome to perform were then assigned to the three aforementioned MSC groups; for instance, walking and crouching for lower extremity MSC. The 15 ‘physical agility’ EK FCE tests were acknowledged as tests measuring activities that were limited by either upper extremity MSC, back MSC or lower extremity MSC. Finally, as a limited activity could be measured by several EK FCE tests, a selection of three EK FCE tests per MSC group was made with regard to the characteristics of the EK FCE tests.

This three-step procedure was believed to avoid eventual the redundant information resulting from assessing the same limited activity through several EK FCE tests, thereby enhancing practicality to some extent in order to assess physical work-ability in a more specific and efficient way. Then, one isometric EK FCE lifting test, one dynamic EK FCE lifting test and one manipulation EK FCE test for each of the three MSC groups were selected. As two EK FCE tests were relevant for two MSC groups, seven EK FCE tests were finally involved in the first empirical study on reproducibility (Chapter 3; Table 2). With regard to the findings in this thesis, only the five EK FCE lifting tests were found to be reproducible in subjects without MSCs or in subjects suffering from LBP. Therefore, these tests were used in the validity studies (Chapter 5 and Chapter 6), revealing that especially the Carrying lifting strength test was found to be moderately valid for the construction industry in a return to work context.
Clinimetrics

As stated in chapter 1, clinimetrics is derived from psychometrics and has become a fundamental methodological discipline in clinical research and clinical practice over the last decade (9,10). Clinimetrics is a fundamental and ineluctable process, as information provided by any clinical instrument cannot be trusted or used in any judgment or decision making process when the measurement quality (i.e. clinimetrics) has not been positively evaluated. Reflected through reproducibility (i.e. reliability and agreement) and validity, clinimetrics deals with the measurement quality of clinical instruments and is based on classical test theory, relying on the decomposition of an observed score into true and error scores (9-15). Measurement error can be either random and associated with reproducibility or systematic and associated with the aspect of validity. With regard to these theoretical and methodological considerations about clinimetrics, a two-step model was proposed and illustrated in chapter 1 (Figure 2) to assess the measurement quality of the EK FCE tests. Conformably to this two-step model, this thesis reflects our empirical quest for scientific evidence of the measurement quality of the EK FCE tests, reflecting both aspects of reproducibility and validity. To illustrate the clinimetric assessment of the EK FCE tests, three EK FCE tests, the Forward manipulation test, the Back-torso lift test and the Carrying lifting strength test, are taken as in figure 1 in order to fit the results of our studies into the two-step model, the levels of reproducibility and validity being reported concordantly with the methodological quality appraisals of our literature review and being noted as ‘-’ for low, ‘+/-’ for moderate and ‘+’ for high.

The first step in the assessment of the clinimetrics of the Forward manipulation test, the Back-torso lift test and the Carrying lifting strength test lies in the evaluation of reproducibility (Figure 1, nr 1). As described in chapter 3, reliability of these EK FCE tests was assessed in subjects without MSDs; a high level (+) of reliability was found for the Back-torso lift test and Carrying lifting strength test and a poor level (-) of reliability for the Forward manipulation test. Furthermore, a reproducibility (i.e. reliability and agreement) study was conducted in subjects suffering from LBP (Chapter 4) and showed a good reliability and agreement level (+) for the Back-torso lift test and
Carrying lifting strength test. All in all, the Back-torso lift test and Carrying lifting strength test passed the reproducibility assessment while the Forward manipulation test failed, suggesting the presence of random error in its measurements. Consequently, both the Back-torso lift test and Carrying lifting strength test can flow through the following clinimetrics step, while adjustments of the Forward manipulation test should be considered in order to reduce random error, leading eventually to a successful reproducibility assessment.

Figure 1: Empirical quest of clinimetrics of three EK FCE tests (Clst, Carrying lifting strength test; Btlt, Back-torso lift test; Fmt, Forward manipulation test; -, low evidence; +/-, moderate evidence; +, high evidence)
The second step of our clinimetrics model involves the assessment of validity (Figure 1, nr 2). In both of our validity studies, the validity of the Carrying lifting strength test and Back-torso lift test was assessed in construction workers in a return to work context, which should be kept in mind when evoking an eventual implementation of these tests. As described in chapter 5, construct validity of the Back-torso lift test and Carrying lifting strength test was assessed in construction workers on sick leave due to MSDs. Substantial differences between two groups of construction workers (discriminative validity) were found in the expected direction (+), but convergent validity with self-reported pain and disability was not established (-). Criterion-related validity (Chapter 6) was evaluated in the construction industry through the assessment of concurrent validity with future work disability risk and through the assessment of predictive validity on future work disability risk and full durable return to work. The Back-torso lift test showed poor levels of concurrent and predictive validity (-), while the Carrying lifting strength test showed moderate evidence of concurrent and predictive validity (+/-). Accordingly, the Back-torso lift test failed the validity assessment, suggesting the presence of systematic error in its measurements. Adjustments should be made and clinimetrics should be reassessed. Unlike the Back-torso lift test, the Carrying lifting strength test passed (but only moderately) the validity assessment.

These findings appear to be limitations of the EK FCE method, as six out of the seven evaluated tests were found to be either not reproducible or not valid for predicting future work disability risk or return to work. Of the seven EK FCE tests evaluated, only the Carrying lifting strength test remains for eventual implementation in occupational medicine for return to work decision making, especially in health and safety services for jobs involving manual material handling. However, as only moderate scientific evidence on the measurement quality of the Carrying lifting strength test is available (Figure 1, nr 3), implementation of this test alone appears at the moment doubtful (Figure 1, nr 4), and its added value for occupational professionals should be studied through empirical research. Nevertheless, it can still be cautiously suggested that reproducible and valid EK FCE tests could be involved in the estimation of current, and only current, physical work-(in)ability, providing occupational professionals with information about someone’s (in)ability to perform an activity in the short term or at this moment. Several studies
have shown that tests from different FCE methods devised to measure the same activity produced different results because of differences in terms of the materials needed and test procedures, underlying the lack of consensus concerning FCE’s operationalization (6,16-18). As FCE tests from different FCEs cannot be used interchangeably, the results of this thesis on the reproducibility and validity of EK FCE tests do not appear generalizable to similar tests from other FCE methods.

Methodological considerations

With regard to the research questions formulated in this thesis and to the appreciation of the main findings already described in this chapter, it seems imperative to summarize and discuss some methodological and procedural features. In the next section, the following issues are discussed: the study populations used and the response rate in the longitudinal study, the designs and procedures used in the various studies, the statistical analyses applied, and finally, the idea of a gold standard versus a reference test.

Study populations and response rates

In the reproducibility studies, the objective was to reduce as much as possible the variation in test scores due to the study population in order to quantify optimally the variation due to either raters or random error. Therefore, because medical conditions might vary between days and are likely to influence performance in repeated measurements, it is acceptable to assume that subjects without MSCs show less variation in performance or test scores in repeated measurements than subjects suffering from MSCs. From this perspective, our choice to assess the reproducibility of the EK FCE tests first in subjects without MSCs and second in subjects suffering from LBP (Chapter 3 and Chapter 4) seems to be grounded.
A strength of both validity studies presented in this thesis (Chapter 5 and Chapter 6) concerns their assessment of construction workers in a return to work context, i.e. occupational medicine. Most of the workers in the construction industry are exposed to high physical work demands, which are strongly related to the occurrence of MSCs and MSDs and to sick leave (19,20). Besides, among all manual material handling activities performed in the jobs of the construction industry, lifting is a particularly dominant activity, making our choice of the use of construction workers in a return to work context for the assessment of the EK FCE lifting tests validity even more relevant. In the criterion-related validity study, only a limited number of drop outs were observed (16%) during the one year follow-up period, despite the fact that almost all participants (85%) on sick leave at baseline did return to work within this one year follow-up period. Seventy-two participants were assessed at baseline in order to retain 60 participants in the study during the whole one year follow-up period. Of the 72 construction workers who were included at baseline, 12 did not participate in one of the assessments either six months or one year later. The main reasons given by the construction workers for dropping out were that they did not have the time or motivation to be assessed again on the EK FCE lifting tests because they had already returned to work, or that they suffered from an MSD that did not allow them to be assessed again on the EK FCE lifting tests because of the exclusion criteria of the EK FCE and our study time table. Sixty construction workers remained over the whole one year follow-up period, which seems to be a reasonable amount, as our original goal was to retain 50 participants for the last assessment of the follow-up period.

**Designs and procedures**
In the evaluation of reproducibility, the time interval between repeated measurements is an issue that requires further comments. Within-subjects designs have the major disadvantage that they have the potential to be confounded either by carry-over effects (i.e. physiological changes) or by practice effects (i.e. learning). Carry-over or learning effects may appear when subjects are influenced by or adapt to the repetition of a test over and over, resulting in variation in test scores not because of the poor reproducibility level of the instrument evaluated, but because of systematic within-
subject variation. Both carry-over and practice effects are directly related to the time interval between repeated measurements.

There is no “gold standard” for the time interval between repeated measurements, and in the literature, authors disagree over whether the time interval should be a few days or a few weeks in order to minimize any physiological or psychological maturation in the subjects (21,22). For instance, the time intervals used in the reproducibility studies retrieved for our systematic review (Chapter 2) ranged from one day to two months (23-27). A good time interval between repeated measurements should recreate in the study population the conditions of the first test moment for the re-administration(s) of the test. Such a “steady state” should minimize the variation in test scores within the subjects in order to prevent, from one test moment to another, any carry-over effects or relevant changes in the health status of the subjects. As a good reproducibility level was found for the EK FCE lifting tests, it can be assumed that the time intervals of four days and three days chosen for our two reproducibility studies appeared well-suited to guarantee a steady state in our study populations. Indeed, we noticed post-hoc that no substantial increase or decrease in EK FCE test scores appeared from the first assessment moment to the second (for instance, a mean test-retest difference of 0.1 kg for the Carrying lifting strength test in subjects suffering from LBP), confirming that carry-over effects were not of importance in our studies. A strength of the design used in our first empirical study (Chapter 3) was the use of three assessment moments, allowing a comparison of interrater reliability over two time intervals (4-day and 8-day time intervals). The results of this study showed that the level of interrater reliability was lower with an 8-day time interval (-0.27 ≤ ICC ≤ 0.90) than with a 4-day time interval (0.74 ≤ ICC ≤ 0.96) in six out of the seven EK FCE tests evaluated, indicating that a shorter time interval might be more appropriate to avoid within-subject variation in test scores from one test moment to another. Therefore, a 3-day time interval was used in the following reproducibility study, and again, minimal within-subject variation was observed. In addition, a counterbalanced within-subjects design was used in both of our reproducibility studies in order to avoid bias due to the importance of rater testing order.
With regard to the reliability of the forward and lower manipulation EK FCE tests (Chapter 3), one procedural aspect could have improved upon afterwards. The manipulation EK FCE tests are functional tests in which the subject’s learning capacity influences considerably the variation of test scores in repeated measurements. For both manipulation EK FCE tests, we found moderate to low levels of reliability because of the large variation in subject test scores from the first to the second assessment period. This variation could be attributed to practice, i.e., learning effects, influencing subjects’ performance at the readministrations of these tests. In order to minimize these carry-over effects due to learning, one or more training session(s) could have been involved in the study procedure before completion of the first assessment of the forward and lower manipulation EK FCE tests. However, as the forward and lower manipulation EK FCE tests were not found to be reproducible in adults without MSDs, neither of these tests were used in our ensuing studies.

It is worth noting that most relevant forms of validity for functional tests (i.e., discriminative, convergent, concurrent and predictive) have been explored in the present thesis. To particularly explore and establish relationships between independent variable(s) at baseline and dependent variable(s) during a follow-up period, an observational prospective longitudinal study design appears as the best suited, even if observational studies provide weaker empirical evidence than do experimental studies (28). To evaluate the concurrent and predictive validity of the EK FCE lifting tests in construction workers (Chapter 6), an (observational) prospective longitudinal study design was used. A strength of this design was the capability of assessing concurrent validity between the EK FCE lifting tests and our reference test, the IDR, at three time points within one year, allowing a comparison over time of degrees of concurrent validity. For instance, the degree of concurrent validity of the dynamic EK FCE lifting tests with future work disability risk changed substantially between baseline (-0.17 ≤ r ≤ -0.12) and the second (-0.47 ≤ r ≤ -0.36) and third (-0.33 ≤ r ≤ -0.23) assessment moments, suggesting that the level of concurrent validity is not stable over time and questioning the durability of this form of validity, especially in a ‘changing’ population recovering from MSDs and sickness absence.
**Statistical analyses**

In recent years, the intraclass correlation coefficient (ICC) has been preferred over the (Pearson) correlation coefficient to express levels of reliability, as it takes into account the agreement within subjects between two test outcomes. Furthermore, ICC allows estimation of the reliability of either a single rater (intrarater or intraobserver) or any number of raters (interrater or interobserver). Furthermore, with the calculation of variance between subjects and between raters and variance due to measurement error, raw agreement in test scores could be calculated and expressed as Standard Error of Measurement (SEM) (Chapter 4). The SEM gives useful information concerning the minimal change in test scores that must be detected in order to attribute this change to a clinically relevant change and not to random error in the population that has been studied. The calculation of SEM is a form of raw agreement that is simple, intuitive and clinically meaningful. However, if no similar data have been previously reported in the literature, as for the EK FCE tests, the establishment of a cut-off value for clinical relevance can not be based on existing evidence and should be based on the practitioner’s knowledge or experience with the evaluated test(s). In our study, SEMs ranged from 1.9 (Upper lifting strength test) to 8.6 (Back-torso lift test) kilograms, appearing especially acceptable for the three dynamic EK FCE lifting tests when comparing their SEMs (1.9 to 3.7 kg) to their means (17.0 to 24.5 kg) and standard deviations (6.3 to 11.2 kg).

In this thesis, several statistical tools have been used to assess the validity of five EK FCE lifting tests. For concurrent and predictive validity, associations were evaluated using Pearson correlation coefficients, and proportions of variance were calculated. For predictive validity, regression analyses are often used to find a model of the independent variables responsible for the prediction of a dependent variable. In our study, it has never been of interest to seek any model within the EK FCE lifting tests or to select some FCE tests that would together make up a good predictive model. We wondered whether the EK FCE lifting tests, independent from each other, could predict our dependent variable. Therefore, no regression analyses were performed. In addition, the calculation of the proportion of variance from the correlation coefficients is the same as the R squared available through regression analyses.
Moreover, in addition to a continuous outcome (percentage), the IDR also gives a binominal outcome based on the cut-off point of 38% set by experts for a higher risk for work disability (29,30). Then, the area under the receiver operating characteristic curve was used in order to evaluate the ability of the EK FCE lifting tests to predict the outcomes of the IDR, which is a well accepted statistical tool for that purpose (31). As we had information on the registered number of days on sick leave until full durable RTW, Cox proportional hazards regression analysis was conducted separately for each EK FCE lifting test, Cox regression being especially adapted and advised to analyse the predictive power of covariates on a time-to-event dependent variable (32). All in all, a broad battery of statistical tools has been used, representing a strength of this criterion-related validity study.

Gold standard versus reference test
As noticed in earlier chapters of this thesis, no gold standard is likely to be available for the assessment of physical work-ability in an occupational setting, either for rehabilitation, return to work or disability claims (33,34). In recent years, authors have suggested different alternatives for the assessment of criterion-related validity when gold standards are missing or imperfect, and terms such as “silver standard” or “reference test” made their appearance in the fields of psychometrics and clinimetrics (35-38). A reference test should act as a second best-case scenario striving to be a well-grounded instrument, commonly used in practice and accepted by experts for a particular setting or sector, and should preferably have some clinimetric evidence.

Due to the unavailability of a gold standard, we have tried to assess the criterion-related validity of the EK FCE lifting tests. We searched for an instrument that could act as a reference test by measuring an affiliated concept as close as possible to the criterion measured by the EK FCE lifting tests. Originally, the EK FCE assess physical work-ability, and the EK FCE lifting tests measure lifting capacity in particular but also reflect other activities relevant for physical work-ability, such as gripping, bending, reaching and walking. Consequently, they also reveal the inability or incapability to perform these activities.
Accordingly, an instrument was sought that would measure an associated relevant concept for the EK FCE lifting tests, dealing with physical work-ability or reflecting the inability to perform physical work. The IDR was then selected to act as a reference test in our criterion-related validity study. The IDR is a well-grounded, widely used and accepted instrument in the construction industry that measures through nine questions the risk for future (physical) work disability due to MSDs (29,30). Furthermore, lifting is one of the most important parameters in occupations of the construction industry, and the nine questions of the IDR addressing physical work-ability indirectly reflect the respondent’s lifting ability. Moreover, jobs in the construction industry are particularly exposed to manual material handling, making this sector relevant for physical work-ability and FCE tests. Therefore, the IDR was chosen to assess the criterion-related validity of the EK FCE lifting tests.

In addition, we also used RTW as a criterion to assess the predictive validity of the EK FCE lifting tests. RTW, reflected in our study as the number of days on sick leave until full durable (for at least four weeks) return to work, is an important outcome in occupational medicine. Whether the term of concurrent validity can still be used when a (self-reported) reference test is used instead of a gold standard remains in doubt, and whether future work disability risk or RTW are affiliated criteria for the EK FCE lifting tests can be argued, but validating a test when no gold standard is available requires gathering evidence of different forms of validity (38).

**General conclusion**

The objective of this thesis was to gather scientific evidence on the clinimetric properties of Functional Capacity Evaluation (FCE) tests, especially the Ergo-Kit FCE tests. Our literature study showed that there was not enough scientific evidence available on the clinimetric properties of most FCEs and that more studies were needed to demonstrate their reproducibility and validity.
With regard to the reproducibility of the EK FCE tests, it can be concluded that two isometric and three dynamic EK FCE lifting tests have moderate to high levels of reliability in subjects without MSCs, while the intra- and interrater reliabilities of the two EK FCE manipulation tests were found to be low. Furthermore, the reliability and agreement of two isometric and three dynamic EK FCE lifting were good in subjects suffering from LBP.

With regard to construct validity, it can be concluded that the construct validity of the EK FCE lifting tests was not supported in construction workers on sick leave due to MSDs. Discriminative validity was not statistically established between a group with a high risk for future work disability and a low-risk group, even if differences in test scores between the two IDR groups occurred in the expected direction, and a poor level of convergent validity was found between the five EK FCE lifting tests and self-reported pain intensity and disability.

With regard to the criterion-related validity assessed in construction workers, it can be concluded that the isometric EK FCE lifting tests have poor levels of concurrent and predictive validity with future work disability risk. The dynamic EK FCE lifting tests showed some moderate evidence of concurrent and predictive validity with future work disability risk in the construction industry, especially the Carrying lifting strength test. Predictive validity of the five EK FCE lifting tests on full durable RTW was not supported, although it was weakly statistically significant for two EK FCE lifting tests.

**Implications for practice**

In the present thesis, scientific evidence on the clinimetrics, *i.e.* reproducibility and validity, of the EK FCE tests has been gathered, either through literature or empirical research, reproducibility being evaluated through repeated measurements and validity studies being conducted in the context of occupational medicine for return to work. In light of the main findings reported herein, some implications can be suggested.
The EK certified raters (users) should be aware that the EK FCE manipulation tests were not reproducible in subjects without MSCs and that the EK FCE lifting tests were reproducible in subjects without MSDs and in subjects suffering from LBP. Certified EK FCE raters working in a rehabilitation context should take these results into consideration and apply only the reproducible EK FCE lifting tests for their evaluation of a patient’s status, \textit{i.e.} for the evolution of his or her physical work-(dis)ability over time, in order to better value clinical changes over time. The EK provider should take these results into consideration and adjust the manipulation tests or procedures in order to enhance their reproducibility.

Originally, the EK FCE strives to report the generic physical work-ability of subjects with or without MSDs. The Carrying lifting strength test alone cannot be assessed in order to gain insight into generic physical work-ability, but rather it provides a measure of a more specific physical work-ability at a given moment for employees working in jobs involving manual material handling.

Just before an employee’s sixth week of sick leave, occupational professionals working in the Dutch return to work context should be able to refer employees sick-listed because of MSDs to the Carrying lifting strength test, especially when they perform jobs involving manual material handling, such as construction workers, firefighters and garbage collectors.

Occupational professionals should also have knowledge of the dynamic EK FCE lifting tests’ interpretation in order to optimally use information from these tests in combination, for instance, with information provided by clinical examination and anamnesis, in their decision-making process for return to work. However, whether information provided by these EK FCE lifting tests has added value for the occupational professionals must first be evaluated.

Occupational professionals working for the Dutch construction industry have use of several instruments, such as the Instrument for Disability Risk and work-ability-graphs (39), for their assessment of (physical) work-(dis)ability. Besides these instruments,
occupational physicians could also incorporate information from reproducible and valid performance-based EK FCE lifting tests in combination, for instance, with information from anamnesis and clinical examination, in their judgment and decision-making process for return to work.

**Future research**

FCEs’ validity should be particularly subject to further research involving different study populations with regard to type of MSDs and occupational contexts in order to generalize its application. In this thesis, the validity of the EK FCE tests was evaluated in a return to work context for the construction industry, as it appeared to be a relevant sector for FCEs. Other sectors where employees are also exposed to manual material handling, *i.e.* heavy physical work demands, should be involved. Furthermore, the discriminative validity of the EK FCE tests has been evaluated within two groups of construction workers and could also focus on two other groups, for instance sick-listed versus not, or either different types of MSDs or different functions or jobs.

The added value of the EK FCE tests has been evaluated in the context of work disability claims (40,41) and should be assessed now in rehabilitation and return to work contexts as well. Particularly in rehabilitation, the aspect of responsiveness is relevant and should also be evaluated through prospective longitudinal designs in order to determine whether FCE tests can provide insight into relevant clinical changes in patient’s status over time as part of a rehabilitation program.

Particular attention should be paid to FCEs’ practicality. As noted in chapter 7, practicality is a critical issue for FCEs and appears limited. Research on a short-form FCE protocol, striving to be less invasive, less generic and less time consuming, should be conducted, for instance for a specific type of MSDs or a particular job. Then, what is the validity of such a short-form FCE protocol?
Finally, studies are needed to support the hypothesis that a combination of information provided from different sources, such as the clinical examination, self-reported questionnaires and performance-based testing, could give occupational professionals a better insight into physical work-(dis)ability, hopefully leading to optimal judgment and decision making in a return to work context.
References

5. Innes E, Straker L. Attributes of excellence in work-related assessments. Work. 2003;20:63-76
40. 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
41. Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. The effect of Functional Capacity Evaluation assessment on the judgment of insurance physicians of the physical work-ability in the context of disability claims. Submitted