Assessment of physical work ability:

the utility of Functional Capacity Evaluation for

insurance physicians

The studies in this thesis were carried out at the Academic Medical Center, Universiteit van Amsterdam, Department: Coronel Institute of Occupational Health, Amsterdam, the Netherlands

Cover design: Rudi Jonker, Redcat productions

Printing: Ponsen & Looijen bv, Wageningen

ISBN: 978-90-9022473-2

Assessment of physical work ability: the utility of Functional Capacity Evaluation for insurance physicians

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof.dr. D.C. van den Boom

ten overstaan van een door het college voor promoties ingestelde

commissie, in het openbaar te verdedigen in de Aula der Universiteit

op woensdag 19 december 2007, te 10.00 uur

door Haije Wind

geboren te Tjilatjap, Indonesië

Promotiecommissie:

Promotor(es):	Prof.dr M.H.W. Frings-Dresen
Co-promotor(es):	dr P.P.F.M. Kuijer dr J.K. Sluiter
Overige leden:	Prof.dr J.W. Groothoff
	Prof.dr J. Dekker
	Prof.dr F. Nollet
	Prof.dr E. Schadé
	Prof.dr J.C.J.M. de Haes

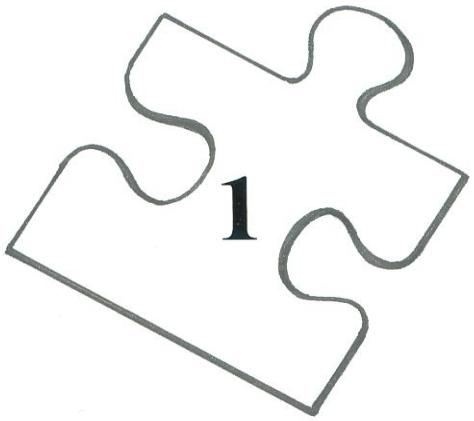
Faculteit der Geneeskunde

Voor mijn vader

Contents:

Chapter 1

General Introduction



1.1 Introduction

At first sight the title of this thesis would seem to have a perfectly clear and obvious meaning, looking at it more closely, several questions arise. What is work ability, and what makes physical work ability such a special theme? What does functional capacity evaluation (FCE) involve? What is utility, and when is an instrument thought to be useful? These different terms will be explained in the following sections, leading to the main research question posed in this thesis. However, in the first place, the special position of insurance physicians (IPs) in the context of this thesis should be clarified. IPs play a role in assessing the level of the employee's work ability in the context of social legislation. In the Netherlands, employer and employee are jointly responsible for arranging the return to work during the first two years of sick leave. After these two years, a disabled worker may claim a disability benefit. It is the statutory responsibility of the IP to assess and record the claimant's work ability, i.e. the extent to which he or she can still carry out certain types of work and the limitations on the work that can be performed. This assessment procedure is subject to rules where consistency, reproducibility and a logical coherence between complaints, disorder, restriction in activities and participation are key concepts. This has consequences for the method IPs use in assessments of work-ability for disability claims, which will be elucidated in the next section.

Work ability

What is work ability? Illmarinen has defined it as follows: "how good is the worker at present, in the near future, and how able is he or she to do his work with respect to the work demands, health and mental resources?" [1]. This definition makes it clear that work ability is not an isolated issue but is embedded in the context of the balance between work load (or work demands) and work capacity (physical and mental resources). The International Classification of Functioning (ICF) offers a framework in which health and health-related domains can be situated [2]. In the ICF model, functioning is described as the interplay between six different model components: disease, body functions and structures, activities and participation, environmental and personal factors. Physical activities are part of the total sum of activities needed to take part in the work process. With very few exceptions, any job will involve a sizable proportion of physical activities. This underlines the importance of physical work ability assessment, and leads us to ask what kind of process this is. Some light can be thrown on this by consideration of the process of clinical diagnosis, which bears certain resemblances to that of work-ability assessment. Research on the reasoning used in making the clinical diagnosis shows that two key processes are involved here: problem-solving and decision-

making [3]. Problems can be solved either by inductive or by hypothetico-deductive reasoning [4]. In the inductive method, the judgment about the diagnosis is delayed until all the relevant information has been collected and pattern recognition matches the test. The hypothetico-deductive method is based on the formation and testing of hypotheses, because clinical reasoning is based on this method. Most clinicians use the latter method in the diagnostic process. Work ability assessment shares many features with the diagnostic process except, of course, that the target is not a diagnosis of a disease, but judgment of work ability. Both involve the collection and processing of information from and about the patient or claimant. In the clinical setting, the most important steps are generating a hypothesis about the medical condition involved, interpretation of additional information to test the hypothesis, pattern recognition and categorization [4]. The steps involved in 'diagnosing' work ability are very similar. The IP starts by collecting information to test the hypothesis that the claimant possesses no residual work ability, and if this hypothesis is rejected, to determine the level of the residual work ability. This is a process shrouded by uncertainty about the accuracy of the outcome. Uncertainty of outcome is a well-known phenomenon in the diagnostic process. It is linked to the second paradigm, namely the medical decision making. It is related to the fact that clinicians work in a situation of uncertainty about the true state of the patient, just as IP's in disability claim assessments remain in uncertainty about the true work ability of the disabled worker.

Probability is a means of expressing - and reducing - uncertainty [5,6]. The normative rule for this process is Bayes' theorem, which states that the information provided by a test can reduce the uncertainty of the outcome if the specificity and sensitivity of the test are high enough. Although the practical implications of this theorem for the assessment of physical work ability are limited, the concept is noteworthy, because the question of this thesis is whether a test can help to reduce the uncertainty about the outcome when IPs are assessing the physical work ability of disability benefit claimants. The hypothesis here is that the claimants have no work ability at all, and the task is to look for information that can provide grounds for rejecting this hypothesis. The 'diagnostic' process involved in the assessment of physical work ability is represented in Fig. 1. In this figure the two key processes (problem solving and medical decision-making) are pictured as methods to reduce uncertainty in the process of assessing the physical work ability by IPs. The practical application of Bayes' theorem to insurance medicine is limited for a number of reasons: the complexity of the decisions about the level of work ability that have to be made, the difficulty of assessing the 'prior probability',

but most of all, the problems involved in determining the expected utility of the outcome, i.e. the assessment of work ability [7].

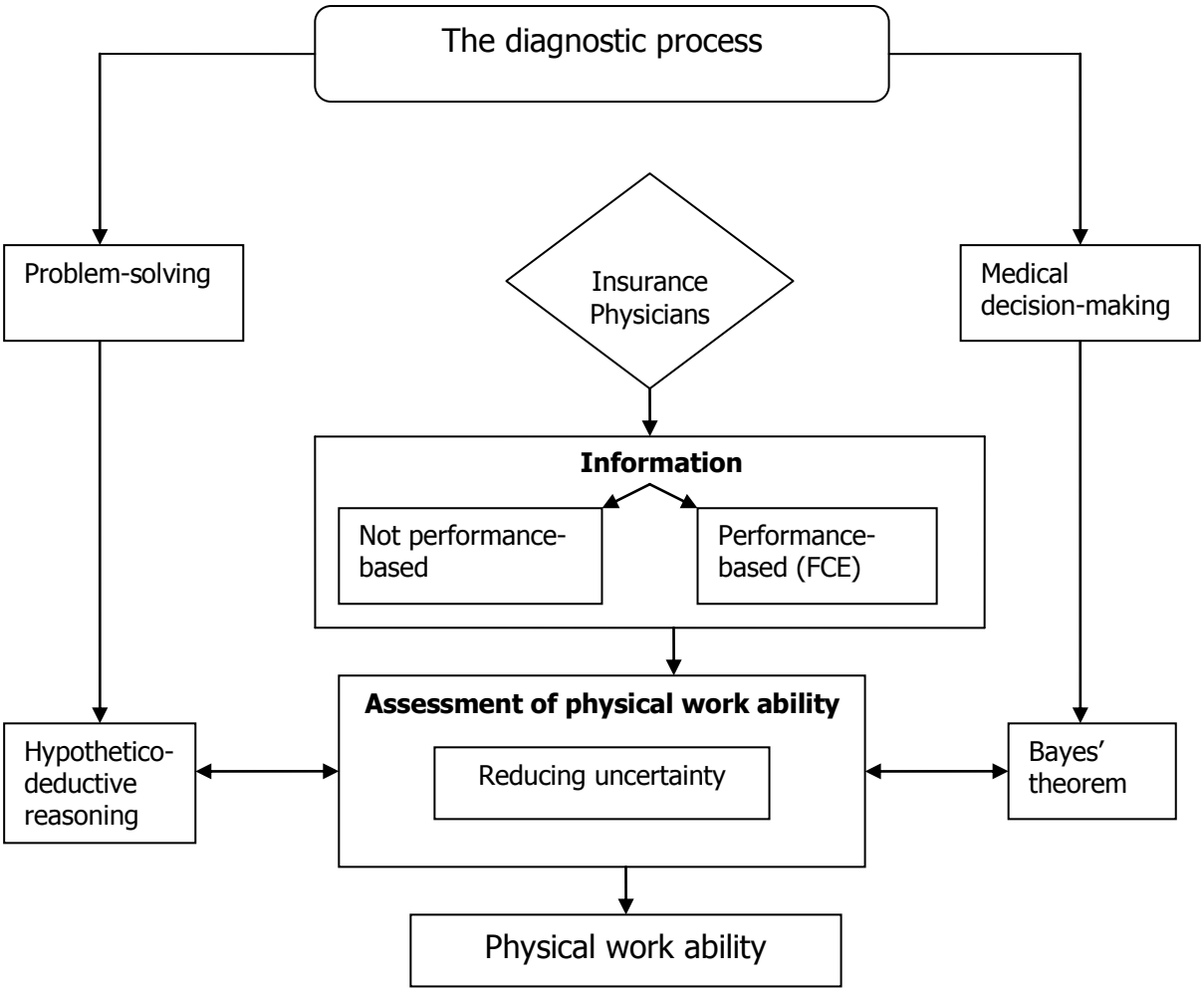


Figure 1: Assessment of physical work ability by insurance physicians, placed in the context of the diagnostic process

One of the current problems afflicting procedures for the assessment of long-term disability is that we have insufficient evidence to base the decision about the work ability upon. Since the claimant is one of the main sources of information for IPs, several methods of handling this information have been developed in the Netherlands [8]. These methods are focused in particular on the question of how to obtain information about the disabled worker, but fail to explain why this information is important for the assessment of work ability and how it can be

translated into concrete estimates of the ability to work. Terms like consistency, reproducibility, and logical coherence are used to try to approach the true physical work ability in insurance medicine. What is needed is more evidence-based information to convert clinical information into restrictions for work. Medical decision-making and evidence-based medicine are closely related [9]. In insurance medicine, the use of both medical decision making and evidence-based medicine stand at the very beginning. Of recent years, however, the evidence-based approach to insurance medicine has concentrated on developing diagnostic standards for specific medical disorders, including non-specific low back pain, which IPs can use in their assessment of physical work ability [10]. Despite these advances, we are still plagued by uncertainty about the precise overall level of work ability after two years of sickness. Assessment of physical work ability is like solving a jigsaw puzzle. Each additional bit of information brings us closer to completing the picture, but some vital pieces are always missing. This thesis is about the question whether a performance based test, in which measuring performance in work-related activities stands central, can help the IP in completing the puzzle of the assessment of the level of physical work ability (see fig. 1).

Musculoskeletal disorders

Physical work ability is placed central in this thesis. Physical work ability is closely related to the presence of musculoskeletal disorders (MSD), but is not confined to only this category of disorders. The ability to participate in work, irrespective what work, is also dependent upon the ability to perform physical activities.

The prevalence of MSDs is high throughout the world [11], and is growing significantly in both developing and developed countries [12]. Several studies reveal the world-wide scale of the problem. One European study on musculoskeletal pain showed that 60-75% of people who experience such pain constantly (in many cases, daily) find that this has a severe impact on their quality of life, limiting their ability to perform physical activities in the context of work and daily life [13].

In the Netherlands, MSD is the second most frequent cause of disability: more than 200,000 persons or 31% of the total number registered as disabled receive a disability benefit for this reason [14]. IPs are therefore regularly confronted with the task of assessing the physical work ability of claimants with MSD. They do not have many instruments at their disposal to support them in this responsibility. The ones that are available for this purpose will be reviewed in the course of this study. The instrument on which our attention will be particularly focused in this thesis is that known as Functional Capacity Evaluation (FCE).

Functional capacity evaluation

FCE is a comprehensive, objective test battery developed to evaluate a person's ability to perform work-related tasks [15]. Reneman [16] lists three fields where information derived from FCE can prove useful: rehabilitation, occupational medicine and insurance medicine. IPs assess the work ability for the settlement of a workers' disability claim. Interest in the use of FCE has been growing at a modest rate in recent years, as reflected in the number of published studies devoted to this subject [17-23]. Some studies approach the use of FCE from the perspective of the disorder limiting functional performance, such as low back pain or upper extremity disorders [18-25]. Others consider the type of work to be done as a starting point for the use or development of an FCE method [26].

Four FCE methods are used in the Netherlands, the Blankenship FCE, the Ergo-Kit, the Ergos Work Simulator and Isernhagen Work Systems. The Blankenship FCE and Ergos Work Simulator make use of a battery of computer-aided tests and require the presence of a qualified rater. The other two FCE methods require the necessary tests to be carried out by a qualified rater. In the context of the present thesis, FCE assessments are performed with the aid of the Ergo-Kit FCE (EK FCE). The reason why we have chosen to perform the study by using the EK FCE is the availability throughout the Netherlands. This makes it possible to execute this study nationwide in the normal procedure of disability claim assessments. There is always an EK FCE facility in the vicinity of the office where the statutory disability claim assessments take place. The EK FCE comprises 55 tests, the complete test protocol lasts about four hours. The tests are based on work-related activities with the following main characteristics:

- Work performed in specific postures (stand, sit, kneel, bend, work above shoulder height)
- Performance of specific activities (walk, lift, carry, crouch, reach, turn, walk up and down stairs, perform short cyclic movements)
- Hand and finger dexterity

As any instrument likely to be used within a diagnostic process, like the assessment of physical work ability, EK FCE should be evaluated with regards to the following five criteria: safety, reproducibility, validity, utility, and practicality. The first criterion to examine is safety. The safety of EK FCE is safeguarded in test procedures, materials used, and rules about exclusion of patients with certain disorders. The test procedures are standardized with rules about the levels to which the persons may be tested. These levels are supervised by trained and certified test raters. Reproducibility (reliability and agreement) and validity, also known as clinimetric properties, refer to the measurement quality of an instrument. A search for evidence about the

reproducibility of FCE, both in the literature and through empirical data, will be performed in this thesis. Concerning validity, there is sufficient proof of the face validity of the Ergo-Kit FCE, considering that the test procedures are standardized and fully described in the user manual. Besides, the procedure of drawing up a report is specified. There is also some proof of content validity of the EK FCE: activities of the test are derived from activities mentioned in the Dictionary of Occupational Titles (DOT)[27]. Evidence for validity of FCE will also be part of a literature study in this thesis. The next criterion, utility of the EK FCE, will be the main theme in the following chapters. Being studied from the perspective of the user of FCE information, this thesis focuses especially on aspects of utility and complementary value of EK FCE information for IPs who might use EK FCE information in the diagnostic process of assessing the physical work ability for disability claims.

What distinguishes FCE from other instruments in disability claim assessments is that it allows the ability to perform specific activities to be assessed under work-related conditions. This is in contrast to non-performance-based methods like anamnesis, X-ray diagnosis and blood tests. While an instrument may provide information that is useful in the assessment of physical work ability, its utility in practice will also depend on the readiness of IPs to accept it. The various aspects of the utility of FCE information derived from Ergo-Kit tests for the assessment of physical work ability in the context of the statutory handling of long-term disability claims for claimants with MSD form the main topic of this thesis. First, however, we need to know what is meant by 'utility' in the context of this thesis.

Utility

The utility of an assessment instrument is directly related to its purpose. An instrument can only be useful if the results obtained with its aid can be used for the planned intervention [28].

The utility of an instrument can be considered at three different levels. The first is that of the organization. At this level, an instrument is considered to be useful when the information it provides helps in achievement of the organization's goals or gives an insight into the quality of the organization's products [29]. The second level is that of the individual user. Seen from this perspective, the information provided by the instrument is useful when it reveals facts hitherto unknown to the user or provides a firmer basis for decision-making about known facts. Moreover, the utility of the instrument also depends upon the frequency with which the instrument can be used and the importance of the information it provides. The third level concerns the intrinsic utility of the instrument itself: is the instrument well designed to meet its purpose [30]? In this thesis, we will be considering the utility of FCE at the second level, that of

the individual user, by studying how IPs can use FCE information to support their assessment of the physical work ability of disability benefit claimants with MSD. FCE is useful in this context when it provides the IPs with information they did not or partly have before or reinforces their judgment as to the validity of the disability claims - i.e., as mentioned above, reduces the uncertainty of the outcome in the IPs' decision-making process.

1.2 Research questions

The results of FCE with the aid of the EK FCE in the context of disability claim assessment are examined in this thesis. The main question posed is: What is the utility of FCE for the assessment of the physical work ability of a claimant with MSD by an IP in the context of statutory long-term disability claim assessments?

This question can be broken down into the following six sub questions:

- What methods are used to assess the physical capacity of the musculoskeletal system in the context of work, daily activities and sport, and what are the reliability and validity of these assessment methods?
- What is known about the reliability and validity of FCE methods?
- What is the reproducibility (i.e. reliability and agreement between raters) of Ergo-Kit tests in subjects with musculoskeletal complaints?
- How do experts in this field perceive the utility of FCE for their work and what arguments do they present to describe the utility of FCE?
- Does information derived from FCE tests lead an IP to change his assessment of the physical work ability of a disability benefit claimant with MSD?
- Is information derived from FCE tests of complementary value to IPs in their assessment of the physical work ability of disability benefit claimants with MSD?

1.3 Hypothesis

On the basis of the research questions stated above, the hypothesis to be tested in this thesis is that IPs consider information derived from FCE tests to be useful as a source of complementary information for the assessment of the physical work ability of long-term disability benefit claimants with MSD.

1.4 Outline of the thesis

In Chapter 2 a systematic review is described of the instruments used to assess the physical capacity of the musculoskeletal system in the context of work, daily activities and sport. The

reliability and validity of these instruments are also described. In Chapter 3 a systematic review is presented of the studies on reliability and validity of several FCE methods, including the EK FCE. In Chapter 4 the reliability and agreement between raters of EK lifting tests is studied in subjects with musculoskeletal complaints. Chapter 5 is devoted to an expert poll in which the utility of FCE as perceived by experts, viz. return-to-work case managers and disability claim experts was studied. Chapter 6 describes a pre/post-test controlled experimental study performed to examine the effect of information derived from FCE tests upon the judgment of IPs in the context of disability claims. The study is based on measurement of the changes in an IP's judgment of the physical work ability of a claimant with MSD in repeated assessments, with and without provision of FCE information between the two assessments. Chapter 7 describes a study of the perceived value of FCE information for the judgment of the physical work ability of disability benefit claimants with MSD by the same group of IPs as that considered in Chapter 6. The IPs were asked whether they regarded FCE information as having complementary value for their judgment of the physical work ability of claimants with MSD, whether provision of FCE information actually led them to change their assessment of the claimants' ability to perform specific work-related activities and whether they would make use of FCE information in future. Finally, the main question of whether FCE tests provide useful information for an IP in the assessment of the physical work ability of claimants with MSD is addressed in the general discussion in Chapter 8, where this issue is placed in the wider context of the assessment of physical work ability as required in the statutory settlement of disability benefit claims.

Reference List

1. Illmarinen J, Tuomi K, Seitsamo J (2005) New dimensions of work ability. International Congress Series. 2005; 1280:3-7
2. International Congress Series. 1280: 3-7WHO (2001) Internation Classification of Functioning, Disabiliy and Health: ICF. Geneve
3. Elstein AS, Schwartz A (2002) Clinical problem solving and diagnostic decision making: a selective review of the cognitive research literature. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ books; pp 179-195
4. Fraser RC (1987) The diagnostic process. In: Fraser RC, ed. Clinical method; A general approach. Leicester: Butterworth; Heineman; pp 35-58
5. Feinbloom RI (1985) The probabilistic paradigm as the basis science of the practice of family medicine. In: Sheldon M, Brooke J, Rector A, eds. Decision- making in general practice. New York: Stockton press; pp 161-166
6. Sox HC, Blatt MA, Higgins MC, Marton K (1988) Quantifying probability. In: Sox HC, Blatt MA, Higgins MC, Marton K, eds. Medical decision making. Boston; London; Durban; Singapore; Sydney; Toronto; Wellington: Butterworth Publishers; pp 61
7. Razenberg PPA (1992) Formation of judgment: scientific framework [Oordeelsvorming: wetenschappelijk kader: in Dutch]. Thesis Universiteit van Amsterdam, Amsterdam; pp 12
8. De Boer WEL, Wijers JHL, Spanjer J, Van der Beijl I, Zuidam W, Venema A (2006) Discussion models in insurance medicine [in Dutch: Gespreksmodellen in de verzekeringsgeneeskunde]. Tijdschr Bedr Verz Geneesk 14 (1): 17-23
9. Elstein AS (2004) On the origins and development of evidence-based medicine and medical decision making. Inflamm res 53: S184-S189
10. The Health Council of the Netherlands (2005) Insurance physicians protocol: Non specific low back pain [Gezondheidsraad. Verzekeringsgeneeskundig protocol Aspecifieke lage rugpijn. Rapport nr. 2005/15: in Dutch]
11. World Health Organization (2003) The burden of musculoskeletal disorders at the start of the new millenium. WHO Technical Report Series 919, 1-218. Geneva, World Health Organization
12. Brooks PM (2006) The burden of musculoskeletal disease - a global perspective. Clin Rheumatol 25: 778-781

13. Woolf AD, Zeidler H, Haglund U, Carr AJ, Chaussade S, Cucinotta D, Veale DJ, Martin-Mola E (2004) Musculoskeletal pain in Europe: its impact and a comparison of population and medical perceptions of treatment in eight European countries. Ann Rheum Dis 63: 342-347
14. Statistics Netherlands (2004) http//www.cbs.nl/ theme labour, income and social security
15. Hart DL, Isernhagen SJ, Matheson LN (1993) Guidelines for functional capacity evaluation of people with medical conditions. J Orthop Sports Phys Ther 18: 682-686
16. Reneman MF, Wittink H (2006) Functional performance evaluation In: Nordin M, Pope M, Andersson M, eds. Musculoskeletal disorders in the workplace II - the prevention of disability
17. Gross DP, Battié MC (2002) Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther 82: 364-371
18. Gross DP, Battié MC (2005) Factors influencing results of Functional Capacity Evaluations in workers compensation claimants with low back pain. Phys Ther 85: 315-322
19. Gross DP, Battié MC (2005) Functional Capacity Evaluation performance does not predict sustained return to work in claimants with chronic back pain. J Occup Rehab 15: 285-294
20. Gross DP, Battié MC, Asante A (2006) Development and Validation of a short-form Functional Capacity Evaluation for use in claimants with low back disorders. J Occup Rehab 16: 53-62
21. Gross DP, Battié MC (2006) Does Functional Capacity Evaluation predict recovery in workers' compensation claimants with upper extremity disorders? Occup Environ Med 63: 404-410
22. Gross DP (2004) Measurement of properties of performance-based assessment of Functional Capacity. J Occup Rehab 14: 165-174
23. Oesch PR, Kool JP, Bachmann S, Devereux J (2006) The influence of a Functional Capacity Evaluation on fitness for work certificates in patients with non-specific chronic low back pain.Work 26: 259-271
24. Soer R, Gerrits EHJ, Reneman MF (2006) Test-retest reliability of a WRULD functional capacity evaluation in healthy adults. Work 26: 273-280

25. Gibson L, Strong J, Wallace A (2005) Functional capacity evaluation as performance measure. Evidence for a new approach for clients with chronic back pain. Clin J Pain 21: 207-215
26. Frings- Dresen MHW, Sluiter JK (2003) Development of a Job-Specific FCE protocol: the work demands of hospital nurses as an example. J Occup Rehab 13: 233-248
27. United States Department of Labor (1991) Dictionary of Occupational Titles, 4th ed., US Government Printing Office. Washington, DC
28. Matheson LN, Mooney V, Grant JE, Leggett S, Kenny K (1996) Standardized evaluation of work capacity. J Back Musculoskelet Rehabil 6: 249-264
29. Van Dijk JFH, De Kort WLAM, Verbeek JHAM (1993) Quality assessment of occupational health services instruments. Occup Med 43: S28-S33
30. Alderson M, McGall D (1999) The Alderson-McGall hand function questionnaire for patients with carpal tunnel syndrome: a pilot evaluation of a future outcome measure. J Hand Ther 12: 313-322

Chapter 2

Assessment of Functional Capacity of the Musculoskeletal System in the context of work, daily living and sport: a systematic review

Haije Wind, Vincent Gouttebarge, P.Paul F.M. Kuijer, Monique H.W. Frings-Dresen
Journal of Occupational Rehabilitation 2005; 15 (2): 253-272

Abstract

The aim of this systematic review was to survey methods to assess the functional capacity of the musculoskeletal system within the context of work, daily activities and sport. The following key words and synonyms were used: functional physical assessment, healthy/ disabled subjects, and instruments. After applying the inclusion criteria on 697 potential studies and a methodological quality appraisal 34 studies were included. A level of reliability > 0.80 and of > 0.60 resp 0.75 and 0.90, dependent of type of validity, was considered high. Four questionnaires (the Oswestry Disability Index, the Pain Disability Index, the Roland-Morris Disability Questionnaire, and the Upper Extremity Functional Scale) have high levels on both validity and reliability. None of the functional tests had a high level of both reliability and validity. A combination of a questionnaire and a functional test would seem to be the best instrument to assess functional capacity of the musculoskeletal system, but need further examined.

2.1 Introduction

We live and so we move. Moving is an important condition to stay healthy [1]. In all sections of our active live, in work, daily activities, sport, the ability to move is important. This ability is strongly related to the function of the musculoskeletal system. A model to register functioning is the ICF (Classification of Functioning, Disability and Health) [2]. Before impairment in functioning can be defined as a restriction, the context must be taken into account. The context defines whether an impairment in moving leads to a restriction in participation and limitation in activities.

Disorders in the ability to move are an important problem in several ways. In relation to work disorders of the musculoskeletal system are important regarding both incidence and costs [3-5]. Musculoskeletal disorders are the most expensive disease category regarding work absenteeism and disablement in the Netherlands [3]. In a study among the Dutch population of 25 years and older, 41% of men and 48% of women reported at least one musculoskeletal disorder in the last 12 months [6]. A high prevalence of musculoskeletal disorders was also found in other countries, such as Great Britain, France, and the United States of America [7,8]. Picavet and Schouten [9] found that more than half of the Dutch population reported low back pain in a period of the last 12 months and almost a quarter of the people with low back pain reported sick leave. Not only the consequences of musculoskeletal disorders in work are important, but work itself is also seen as a major cause of musculoskeletal disorders [10-13]. Work, disability and return to work are closely related concepts. Assessment of functional capacity can not be seen separate from these concepts. Several models for return to work, are nowadays known, but one of the first was proposed by Feuerstein in 1990 [14]. In daily activities and sport decline of the functional capacity of the musculoskeletal system lead to restrictions in participation and limitation in activities [15-18]. In growing older daily activities get restricted and limited by the reduction in mobility, muscle strength and coordination [15]. The relation between sport and injuries of the musculoskeletal system has been established in several studies [19-23].

Because moving and the restriction in moving are so important and have great personal and financial consequences, an accurate assessment of the restriction in participation is important. Currently, there are several ways in which the functional capacity of the musculoskeletal system can be assessed. These assessments are performed by occupational and vocational rehabilitation providers, such as occupational therapists, occupational and rehabilitation physicians, and physiotherapists. The most widely used instruments to assess physical capacity are questionnaires [24-26] and tests [27-29]. Millard [30] presents in a critical review

14 questionnaires of which some assess the functional capacity in the context of work and daily activities. However, to our knowledge, there is no systematic overview that describes the different instruments used and their quality in terms of reliability and validity. Therefore, the research questions of this systematic review are:

- What methods are used to assess the functional capacity of the musculoskeletal system in a specific context?
- What is the reliability and validity of these assessment methods?

2.2 Method

Search strategy

The literature was identified by means of a systematic computerized search of the following bibliographical data bases: Medline (biomedical literature, 1966- October 2003); Embase (biomedical and pharmacological literature, 1980 – October 2003); Cinahl (nursing and allied health, 1982 – October 2003); RILOSH (health and safety at work, 1975- October 2003); MIDHAS (health and safety at work, 1985 – October 2003); HSELINE (health and safety at work, 1987- October 2003); CISDOC (safety and health at work, 1987- October 2003) and NIOSHTIC (workplace safety and health, 1990- October 2003). The following key words were used: functional physical assessment, healthy/ disabled subjects, and instruments. The synonyms are listed in Table I. The synonyms were connected by 'or'. To complete the search strategy we connected the results of each column of synonyms by 'and'.

Selection

Inclusion criteria were defined and used to acquire all relevant literature. In order to be eligible for inclusion a paper had to meet the following criteria:

1. The paper had to be written in English, Dutch, French or German.
2. The paper had to describe the method to assess functional capacity of the musculoskeletal system. Functional capacity of the musculoskeletal system was defined as the physical ability of a subject to perform functional activities.
3. The paper had to describe the context of the assessment: work, daily activities, or sport.
4. The paper had to describe results based on a human population.

Study selection

In Step 1 the first two authors (HW, insurance physician, and VG, human movement scientist) independently reviewed the titles of the studies that were selected on the basis of the key words and their synonyms by applying the inclusion criteria 1 and 4. In Step 2 the abstracts of the remaining studies were read and the inclusion criteria applied. The abstracts that fulfilled the inclusion criteria were included for the full text selection. If the abstract did not provide enough information, according to the reviewers, to decide whether or not the inclusion criteria were met, the study was included for the full text selection. In Step 3, the inclusion criteria were again applied by the same two reviewers independently. Disagreements, if any, on the inclusion or exclusion of articles were resolved by consulting a third reviewer (PK). Review studies were included and only used to screen for more original papers. Furthermore, the selection of papers was extended by screening the reference lists of all selected studies by applying the inclusion criteria. But the reference lists of the papers that were selected from the reference lists of included articles and reviews were not searched for additional studies.

Table 1: The key words and their synonyms used in the literature search

Functional Physical Capacity	Healthy/disabled subjects	Instruments
Functional	Healthy subjects	Investigation
Occupational	Disabled subjects	Interview
Vocational	Musculoskeletal	Questionnaire
Work	Locomotor	Medical examination
Work-related	Limb	Physical examination
Employment	Extremity	Examination
Job	Low back	Anamnese
Physical	Spine	Anamnesis
Career	Spinal	Instrument
Profession	Neck	Measure method
In combination with		Measurement
Assessment		Instrumentation
Evaluation		Scale
Capacity		
Testing		
Simulation		
Performance		
Rehabilitation		

Methodological quality assessment

The selected studies were rated on methodological quality by the two reviewers (HW and VG), independently, on the basis of a standardized set of criteria. Table 2 lists the criteria for the assessment of the methodological quality of the included papers. As the methodological

quality of a study influences the results and conclusions, a three-level quality appraisal scale was developed to evaluate the scientific quality of each study. This scale was based on several studies [30-32].

Table 2: The criteria for the assessment of the methodological quality of the included studies based on several authors [31-33].

Objective of the study
+ the objective is clearly described.
± the objective is indistinct, assigning '+' or ' – ' is not possible
– the objective of the study is missing or essential elements are missing
Design
+ true experimental; quasi experimental and multiple measures
± quasi experimental, single study; non experimental, multiple measures
- non experimental
Population
+ the main features are clearly described including age, gender, and medical status. The sample size is appropriate for the population to which the findings are referred. The source of subjects is evident.
± the description of the main features is indistinct, assigning '+' or ' – ' is not possible.
- the main features of the sampling frame are not described and the population and/ or the sample of subjects is not appropriate to the population to which the findings are to be referred.
Assessment method
+ the assessment method is clearly described. In case of a questionnaire and interview, the questions are comprehensible. In case of an examination, the precise actions are described. In case of a technical device, the measurement procedure is described.
± the assessment method is indistinct, assigning '+' or ' – ' is not possible.
- the assessment method is not described or essential parts are missing.
Analysis and presentation
+ all statistical procedures to analyse are described. The statistical procedures are appropriate and correctly used. The presentation is unambiguous and presented tables and figures support the text.
± the statistical procedures are described, but the procedures are not appropriate and/ of incorrectly used. There are mistakes in the use of the statistical procedures. The presentation is ambiguous.
- the statistical procdures of which the results are described are not mentioned or there is some statement about the use of statistical procedures, but the procedures are inappropriate and incorrect. There are grave mistakes in use of the statistical procedures. The presentation is ambiguous.

The criteria concerned the objective, population, assessment method, the study design, and the analysis and presentation of the statistical outcome. To be admitted to the discussion of this review a study had to have at least three out of the five possible positive appraisals for the abovementioned criteria. Studies that did not meet this standard were not described any further. Disagreements between the two reviewers were subsequently discussed during

consensus meetings. If disagreements could not be solved during such a meeting, the third reviewer (PK) was consulted for a final judgment.

Reliability and validity:

Reliability is the extent to which an experiment, test, or any measuring procedure yields no difference in results of repeated trials. The concept of reliability is a fundamental way to reflect the amount of error, both random and systematic, inherent in any measurement [34]. Error-free measurement can never be obtained [35]. Different types of reliability are known [36]. In our study we judged the following basic methods for estimating the reliability of the instrument: intrarater and interrater reliability, internal consistency and test-retest reliability. The interrater and intrarater were generally expressed as a correlation coefficient. Internal consistency was expressed by the kappa (κ) or Crohnbach's alpha [37] and test-retest by a correlation coefficient, percentage agreement, or the kappa (κ).

Validity is the extent to which an experiment, test, or any measuring procedure measures what it is intended to measure. Just like reliability, validity is also a matter of degree [38]. For validity we rated the following standards for estimating the validity of the instrument: face and content validity, criterion validity, and construct validity. We rated face and content validity as high, moderate and low, depending on the extent to which the test was found to measure what it was supposed to measure and the extent to which it covered all the relevant dimensions and aspects that were supposed to fit in the test [39]. For criterion validity (concurrent and predictive) statistical measures like percentage agreement, correlation and kappa coefficient were used. Construct validity (convergent and discriminant) was expressed as a correlation coefficient. For responsiveness of the instrument, we used a number of standards, such as the correlation between test results preoperative and postoperative, and also pre-treatment and post treatment, the area under the ROC, and effect size. The balance between sensitivity and specificity of a test can be examined using a graphic presentation called a receiver operating characteristic (ROC) curve. The area under the curve is an indication of 'goodness' of the test. A non-discriminating test has an area of 0.5, and a perfect test has an area of 1.0 [40]. The limiting values of the different types of reliability and validity and the appraisal are listed in Table 3. Studies that did not describe the reliability and validity of a test were not described any further. When referred to in former studies, those levels of reliability and validity were used.

Table 3: The levels of reliability [36,97,98], validity [98,99] and responsiveness [100-103] for the methodological quality assessment

Level of reliability: intrarater reliability, interrater reliability and internal consistency, test-retest

Intrarater, interrater reliability
- Pearson Product Moment Coefficient (r), Spearman Correlation Coefficient (p)
  - high: r / p > 0.80
  - moderate: 0.50 < r/ p < 0.80
  - low: r / p < 0.50
- Percentage of agreement %
  - high: % > 0.90 and the raters can choose between more than two score levels
  - moderate: % > 0.90 and the raters can choose between two score levels
  - low: The raters can choose only between two score levels
- Intra-class Correlation Coefficient ICC
  - high: ICC > 0.90
  - moderate: 0.75 < ICC < 0.90
  - low: ICC < 0.75

Internal consistency
- Intra-class Correlation Coefficient ICC
  - high: ICC > 0.90
  - moderate: 0.75 < ICC < 0.90
  - low: ICC < 0.75
- Kappa value k
  - high: k > 0.60
  - moderate: 0.41 < k < 0.60
  - low: k < 0.40
- Cronbach's Alpha α
  - high: α > 0.80
  - moderate: 0.71 < α < 0.80
  - low: α < 0.70

Test-retest
- Pearson Product Moment Coefficient (r), Spearman Correlation Coefficient (p)
  - high: r / p > 0.80
  - moderate: 0.50 < r / p < 0.80
  - low: r / p < 0.50
- Percentage of agreement %
  - high: % > 0.90 and the raters can choose between more than two score levels
  - moderate: % > 0.90 and the raters can choose between two score levels
  - low: The raters can choose only between two score levels
- Kappa value k
  - high: k > 0.60
  - moderate: 0.41 < k < 0.60
  - low: k < 0.40

Level of validity

- Face / Content validity
  - high: The test measures what it is intended to measure and all relevant components are included
  - moderate: The test measures what it is intended to measure but not all relevant components are included
  - low: The test does not measure what it is intended to measure
- Criterion-related validity: concurrent and predictive validity
  - high: Substantial similarity between the test and the criterion measure (percentage agreement ≥ 90%, k > 0.60, r > 0.75)*
  - moderate: Some similarity between the test and the criterion measure (percentage agreement ≥ 70%, k ≥ 0.40, r ≥ 0.50)*
  - low: Little or no similarity between the test and the criterion measure (percentage agreement < 70%, k < 0.40, r < 0.50)*
- Construct validity: convergent and divergent validity
  - high: Good ability to differentiate between groups or interventions, or good convergence / divergence between similar tests (r ≥ 0.60)
  - moderate: Moderate ability to differentiate between groups or interventions, or moderate convergence / divergence between similar tests (r ≥ 0.30)
  - low: Poor ability to differentiate between groups or interventions, or low convergence / divergence between similar tests (r < 0.30)

Level of Responsiveness

Significant difference in T-test:

high	Significant difference ($P \leq 0.05$) between groups over time in scores
low	No significant difference

Area under Receiver Operating Characteristic (ROC) curve:

high	$AUC > 0.75$
moderate	$0.5 \leq AUC \leq 0.75$
low	$AUC < 0.5$

Effect Size:

high	$Es \geq 0.8$
moderate	$0.4 \leq Es < 0.8$
low	$Es < 0.4$

2.3 Results

Literature search

The literature search in the various databases on the key words resulted in a selection of 1227 publications. After removal of duplications, 697 studies remained. The first search on title resulted in exclusion of 42 studies. Thirty-seven studies were not written in English, French, German, or Dutch and five studies had no data based on human subjects. The application of the inclusion criteria to the abstracts eliminated 563 studies. Some studies were excluded on the basis of more than one inclusion criterion. Seven studies appeared not to be based on data of a human population, 393 studies failed to describe the functional relevance. In 184 studies the disorder was not musculoskeletal, and 423 studies described no context. A total of 92 studies remained, and the inclusion criteria were applied to the full text. Of these 92 studies, four studies could not be obtained. Of two studies the publisher could not be found and two studies had no correct references. Forty-six studies were excluded: ten studies did not use a functional assessment method, 28 studies had no context, and in eight studies neither of the criteria was found. As a result, forty-two studies remained: 34 original papers, and eight reviews. All papers and seven reviews were written in English. One review was written in German. Another 14 studies were identified from the screening of the bibliography of these original papers and reviews: nine studies from the reviews and five from the original studies. The present study, therefore, included 48 original articles. Agreement between the two reviewers on the inclusion criteria was nearly perfect (95%). For the remaining studies the third reviewer was consulted to make a final decision.

Methodological quality appraisal

After application of the methodological appraisal, 14 studies [15,41-53] received less than three positive ratings. The level of agreement between reviewers in assessing these appraisals was

excellent (100%). The methodological quality of the remaining 34 studies was sufficient and they are presented.

Studies included

The methods assessing functional capacity of the musculoskeletal system can be divided into questionnaires and functional tests. Thirteen questionnaires and 14 functional tests were described in the different studies. These questionnaires and tests can be divided into methods designed to assess the general functioning and the specific functioning of the musculoskeletal system.

Questionnaires

Two questionnaires [54,55] described general functioning, and 11 questionnaires described specific functioning. Seven questionnaires assessed the functional capacity of the low back [24,25,56-60] and one questionnaire assessed the functional capacity of the neck [61]. Two questionnaires assessed the functional capacity of the upper extremity [62,63] and one [70] questionnaire assessed the functional capacity of the lower extremity. In eight questionnaires the context was work [24,25,56-60,63], in two questionnaires the context was work and daily activities [25,64] and in three questionnaires the context was daily activities [55,61,62]. No questionnaires were found in the context of sport. Although the 11 questionnaires were specific, the authors concluded that the tests could be used for the measurement of general functioning, except the questionnaires for upper and lower extremities. In Table 4 the characteristics of the included questionnaires are presented.

Table 4: Questionnaires to assess the functional capacity of a person and a description of the questionnaires in terms of area (general, specific) activities, type of scale, measurement, context (work, daily activities, sport), study design, and the characteristics of the population

	Area General / Specific:	Activities; Scale Type Scale: R: ratio I: interval O: ordinal N: nominal	Measure-ment	Context W: work S: sport A: daily act.	Study design true experimental quasi experimental non-experimental --- pre-post; post only time series; multiple measures; single study	Population N: Number of subjects A: Age: mean age, range, sd G: Gender H: Health status	Author
Disability Rating Index (DRI) [54]	General	Dress, walk, stairs, sit, stand, carry, household activities, run, light work, heavy work, lift, participate in work, sport I: visual analogue scale	General functioning	W	Quasi-experimental Multiple measures	N: 1092 A: 43 (17-76) G: 567 males; 525 females H: healthy N: 366 A: 50 (21-85) G: 135 male,231 female H: musculoskeletal disorders, multiple sclerosis	Salèn B. A. et al 1994 [54]
					Non-experimental pre-posttest	N: 114 A: 44 G: 46 males; 68 females H: Low back pain	Strand L.I. et al 2002 [80]
Medical Rehabilitation Follow Along (MFRA) [55]	General	Personal care, lift, walk, travel O: 6 levels Get up, stairs, sit, stand, reach, kneel, drive O: 3 levels	General functioning	A	Quasi-experimental Single study	N: 47; A: 46 (19-72) G:18 males; 29 females H: Low back pain carpal tunnel syndrome, other	Granger C.V. et al 1995 [55]
MOS 36-item Short Form Health survey [56]	Specific: Low back	Physical functioning: lift heavy objects, lift groceries, stairs, bend, kneel, stoop, walk, run, move, push O: 3 levels	General functioning	W	Non-experimental multiple measures	N: 6 A: 43 (37-66) G: 5 males; 1 females H: Low back pain	Harwood K.J. 2001[67]
					Non-experimental multiple measures	N: 42 A: 40.2 (8.9) G: 31 males; 11 females H: injury; chronic pain work related	Hart D.L. 1998 [69]
					Non-experimental multiple measures	N: 19 A: 40.5 (24-57) G: 10 males; 9 females H: thoracal or lumbar spine fracture	Leferink V.J.M. et al 2003 [73]
Million Visual Scale [57]	Specific: Low back	Stiffness, walk, stand, turning, twisting, sit, lie, daily tasks, work I: visual analogue scale	General functioning	W	Non-experimental pre-post and multiple measures	N: 1749 A: 41 (10) G: 1102 males; 647 females H: chronically disabling spine disorder	Anagnostis C. et al 2003 [105]
Oswestry Disability Questionnaire (ODQ) [24]	Specific: Low back	Pain, personnel care, lift walk, sit, stand, sleep, sex life, social life, travel O: 6 statements	General functioning	W	Non-experimental pre-posttest	N: 42 A: 38 (17-63) G: 28 males; 14 females H: Low back pain	Di Fabio R.P. et al 1996 [106]
					True- experimental multiple measures	N: 110 A: 40 (22-61) G: 64 males; 48 females H: Back pain	Loisel P. et al 1998 [107]
					Non-experimental multiple measures	N: 18 A: 35.7 ± 7.1 G: 11 males; 7 females H: Low back pain	Parks K.A. et al 2003 [68]
					Non-experimental multiple measures	N: 6 A: 43 (37-66) G: 5 males; 1 females H: Low back pain	Harwood K.J. 2001[67]

Oswestry Disability Questionnaire (ODQ) Continued					Non-experimental multiple measures	N: 42 A: 40.2 (8.9) G: 31 males; 11 females H: injury; chronic pain work related	Hart D.L 1998 [69]
					True experimental time series	N: 111 A: 40.4 (22-61) G: 63 males, 48 females H: low back pain	Poitras S. et al 2000 [108]
Pain Disability Index (PDI) [58]	Specific: Low back	7 areas of daily living: family/ home responsibilities, recreation, social activity, occupation, sexual behaviour, self care, life-support activity O: 10 levels of pain-rating	General functioning	W	Non-experimental multiple measures	N: 42 A: 36.5 (8.5) G. 34 males; 8 females H: Pain related disability	Gibson L. & Strong J 1996 [60]
Questionnaire Physical Activities [59]	Specific: Low back	Sit R: % total time Hands above shoulder, hands below knee, bend and twist, repetitive hand/finger movements, lift, carry O: 5 statements	General functioning	W	Non-experimental multiple measures	N: 484 A: 48.5 G: 232 males; 252 females H: healthy and low back pain	Torgen M. et al 1997 [59] Torgen M. et al 1999 [109]
Roland Morris Disability Questionnaire (RMDQ) [25]	Specific: Low back	24 Activities: among these: walk, work, climb, rest, get up, stand, bend, kneel, pain, turn in bed, dress, sleep, sit, N: yes/no	General functioning	W; A	Non-experimental multiple measures	N: 19 A: 40.5 (24-57) G: 10 males; 9 females H: thoracal or lumbar spine fracture	Leferink V.J.M. et al 2003 [73]
Spinal Function Sort (SFS) [60]	Specific: Low back	50 drawings depicting performance manual material handling tasks (DOT) like lifting, bending, carrying O: 5 statements	General functioning	W	Non-experimental multiple measures	N: 42 A: 36.5 (8.5) G. 34 males; 8 females H: Pain related disability	Gibson L. & Strong J 1996 [60]
Neck Disability Index (NDI) [61]	Specific: Neck	Pain intensity, personal care, lift, sleep, drive, recreation, headache, concentration, read, work. O: 6 statements	General functioning	A	Non-experimental multiple measures	N: 48 A: 37 (18-55) G:17 males, 31 females H: neckpain	Vernon H. & Mior S 1991 [61]
Activities of Daily Living Upper extremity [62]	Specific: Upper extremity	Ambulate, feed, dress, perform personal toilet, can communicate O : 3 grades	Functioning upper extremity	A	Non-experimental Single study	N: 79 A: - (<40 > 90) G: 41 males, 38 females H: hand disorders	Carroll D. 1965 [62]
Upper Extremity Function Scale (UEFS) [63]	Specific: Upper extremity	Sleep, write, open jars, pick up small objects, drive, open door, carry, wash dishes O: 10 degrees	Functioning upper extremity	W	Quasi-experimental multiple measures	N: 108 A: 38 (19-65) G: 36 males; 72 females H: upper extremity disorders N: 91 A: 46 (22-80) G: 30 males 61 females H: CTS patients	Pransky G. et al 1997 [63]
Lower Extremity Activity Profile (LEAP) [64]	Specific: Lower extremity	Self care, mobility, household, leisure I: visual analogue scale	Functioning lower extremity	W; A	Non-experimental pre-posttest	N: 32 A: 66 (SEM 1.2) G: 14 males; 18 females H: knee disorders	Finch E. & Kennedy D. 1995 [64]

Functional tests

Six functional tests [65-71] described general functioning, and eight tests described specific functioning. Of these eight tests, four functional tests [72-75] assessed lift capacity. One test assessed the functional capacity of the hand [28], one test assessed the functional capacity of the upper extremity [62], and two tests assessed the functional capacity of the lower extremity [76-79]. In eight of the functional tests the context was work [29,65-69,72-75]. In four functional tests [28,62,70,71] the context was daily activities and in the two functional tests for the lower extremity

the context was sport [76-79]. For eight tests the authors concluded that the tests could be used for the measurement of general functioning [65-71,73]. The other tests were used to measure the functioning of the area assessed in the test, such as the Jebsen Hand Function Test [28] to measure functioning of the hand and the Functional Performance Tests to measure functioning of the lower extremity [76-79]. Table 5 lists the characteristics of the included functional tests.

Table 5 : Functional tests to assess the functional capacity of a person and a description of the questionnaires in terms of area (general, specific) activities, type of scale, measurement, context (work, daily activities, sport), study design, and the characteristics of the population

	Area General / Specific:	Activities; Scale Type Scale: R: ratio I: interval O: ordinal N: nominal	Measure-ment	Context W: work S: sport A: daily act.	Study design true experimental quasi experimental non-experimental --- pre-post; post only time series; multiple measures; single study	Population N: Number of subjects A: Age: mean age, range, sd G: Gender H: Health status	Author
Baltimore Therapeutic Equipment [65]	General	Wheel turn, push, pull, overhead reach R	Functioning Upper extremity	W	Non-experimental Multiple measures	N: 20 A: 24.8 (18-39) G: 20 males H: healthy	Bhambhani Y. et al 1993 [65]
DOT Residual Functional Capacity [66]	General	Stand, walk, sit, lift, carry, push, pullstop, climb R	General functioning	W	Quasi-experimental Single study	N: 67 A: 41.0 (10.1) G: 37 males, 30 females H: Chronic low back pain	Fishbain D.A. et al 1994 [66]
		crawl, balance, kneel, reach, handle, fingering, feeling shapes N: able/not able	General functioning	W	Quasi-experimental Single study	N: 185 A: - G: - H: Low back pain	Fishbain D.A. et al 1999 [110]
Functional Capacity Evaluation [67]	General	Lift, carry R squat, stand, sit, walk, climb stairs	General functioning	W	Non-experimental Multiple measures	N: 6 A: 41.3 (37-56) G: 5 males, 1 females H: Low back pain	Harwood K.J. 2001 [67]
	General	N: able/not able 5 minutes Handgrip, dynamic pull, lift, carry, walk, sit, stand R	General functioning	W	Non-experimental Single study	N: 18 A: 35.7 ± 7.1 G: 11 males, 7 females H: Low back pain	Parks K.A. et al 2003 [68]
Functional Capacity Evaluation [69]	General	Lift, carry R	General functioning	W	Non-experimental Single study	N: 42 A: 40.2 (8.9) G: 31 males, 11 females D: injury, chronic pain work-related	Hart D.L. 1998 [69]
Physical Performance Tests [70]	General	Pick-up, put on a sock, roll-up O: 3 levels Fingertip-to-floor, lift R	General functioning	A	True-experimental Multiple measures	N: 117 A: 43.8 (10.6) G: 46 males, 71 females H: Low back pain	Strand L.I. et al 2001 [70]
					Non-experimental Pre-posttest	N: 114 A: 43.9 (10.6) G: 46 males, 68 females H: Low back pain	Strand L.I. et al 2002 [80]
Tufts Assessment of Motor Performance (TAMP) [71]	General	Mobility: transfer, sit, rise, stand, walk, walk on ramp, stairs ADL: pour, drink, cut, dress Communication: talk, write, type, paper in envelope O : 4 dimensions; 12 subscales	General functioning	A	Non-experimental Single study	N: 40 A: 25.6 (6-82) G: 14 males, 26 females H: multiple disorders	Gans B.M. et al 1988 [71]

EPIC Lift capacity Test (Employment Potential Improvement Center) [72]	Specific	Lift R	Lift capacity	W	Quasi-experimental Multiple measures	N: 344 A: 30.5 (7.9) G: 168 male, 176 female H: healthy N: 14 A: 31.7 (7.2) G: 9 males, 5 females H: spine, lower extremity impairment	Matheson L.N. et al 1995 [72]
					True experiment pre-post; post only	N: 55 A: 47.2 (12.5) G: 26 males, 29 females H: lumbar spine problems	Matheson L.N. et al 1995 [111]
Lifting tests [73]	Specific	Lift R	General functioning	W	Non-experimental Multiple measures	N: 19 A: 40.5 (24-57) G: 10 males, 9 females H: thoracal and lumber spine fractures	Leferink V.J.M. et al 2003 [73]
Physical Work Capacity [74]	Specific	Lift R	Lift capacity	W	Quasi-experimental Multiple measures	N: 91 A: 26.2 ± 6.5 G: 33 males, 58 females H: healthy	Jackson A.S. et al 1997 [74]
Progressive Isoinertial Lifting Evaluation (PILE) [75]	Specific	Lift R	Lift capacity	W	Non-experimental Multiple measures	N: 160 A: 35.1 (7.5) G: 160 males H: healthy	Mayer T.G. et al 1994 [75]
					Quasi-experimental Multiple measures	N: 22 A: 42 (26-61) G: 22 females H: healthy and various complaints	Horneij E. et al 2002 [29]
Jebsen Hand Function Test [28]	Specific: Hand	Write, turning cards, picking up small objects, simulate feeding, stack checkers, pick up large light objects, pick up large heavy objects R	Hand function	A	Non-experimental Single study	N: 300 A: 20-94 G: 150 males, 150 females H: healthy N: 26 A: 34.5 ± 20 G: - H: hand disorders N: 33 A:- ; G: - H: neurological hand disorders	JebsenR.H. et al 1969[28]
					Quasi-experimental Multiple measures	N: 9 A: 70-78 H: 9 males H: healthy	Chan W.Y.Y. & Chapparo C. 1999 [112]
Upper Extremity Function Test (UEFT) [62]	Specific: Upper extremity	Grasp, grip, lateral prehension, pinch, place, supination and pronation O: 4 levels	Functioning upper extremity	A	Non-experimental Single study	N: 79 A:- (<40 - > 90) G: 41 males, 38 females H: hand disorders	Carroll D. 1965 [62]
Functional Performance Tests [76]	Specific: Lower extremity	Hop 1 leg, triple hop 1 leg, timed hop 1 leg, shuttle run with and without pivot R	Functioning lower extremity	S	Quasi-experimental Time series	N: 93 A: 17-34 G: 58 males; 35 females H: Healthy N: 35 A: 17-34 G: 26 males; 9 females H: knee: ACL deficient	Barber S.D. et al 1991 [76]
		Single hop, triple hop, cross-over hop, timed hop R			Non-experimental Multiple measures	N: 20 A: 24.5 ± 4.2 G: 5 males; 15 females H: healthy	Bolgla L.A.& Keskula D.R. 1997 [77]
		Triple cross-over hop 1 leg shuttle run with pivot R			Quasi-experimental Multiple measures	N: 16 A: 22.9 (18-29) G: 9 males, 7 females H: ankle instability	Munn J. et al 2002 [78]
Motor Activity Score [79]	Specific: Lower extremity	40-m walk, 40-m run, figure 8 run, single hop, cross over hop, stairs hop N: dichotomic	Functioning lower extremity	S	Non-experimental Time series	N: 24 A: - G:- H: ankle sprains	Wilson R.W. et al 1998 [79]

Reliability and validity

The level of reliability of eight questionnaires [24,25,54,57,58,60,61,63] and four functional tests [28,72,75,79] was high. The level of validity was high in six questionnaires [24,25,58,60,61,63] and in one functional test [80]. Responsiveness of three questionnaires appeared from a significant change in the results [24,54,64]. For the Roland-Morris Disability questionnaire, responsiveness based on a ROC curve was moderate to high, depending on the study [81-85]. There were five questionnaires with both high levels of reliability and validity [24,25,60,61,63]. There was no functional test with high levels of both reliability and validity.

A combination of both high reliability and validity testing and extensive validity testing was found in the Pain Disability Index, the Oswestry Disability Index (ODI), and the Roland-Morris Disability Questionnaire (RMDQ) [24,25,58]. Reliability of these questionnaires was high, both on the Intra-class Consistency Correlation and on the test-retest. Validity was also high, especially on construct validity. The selected questionnaires appeared to be responsive to change. The Upper Extremity Function Scale (UEFS) [63] showed both high levels for reliability and criterion-related validity. Among the functional tests, the Back Performance Scale [80] was the test that was most extended studied. Validity was high, but the reliability was moderate. The four questionnaires and the functional test were used in the context of work. Although the aim of three of these questionnaires was to assess the functional capacity of the low back, the authors concluded that the results could also be used to measure general functioning. The UEFS [63] was a questionnaire for assessment of functional capacity of the upper extremities. In table 6 the characteristics of the levels of reliability and validity are presented.

Table 6 : Reliability and validity of the assessment methods

	RELIABILITY				VALIDITY	
Name of assessment method	Interrater correlation	Intrarater correlation	Internal Consistency	Test-retest	Face: F Content: Ct Criterion: (concurrent/predictive) Cr Construct: (convergent/divergent) Co Responsiveness: Re	Author
Questionnaires						
Disability Rating Index (DRI) [54]	*r*: 0.99	*r*: 0.99	α: 0.84	R: 0.95 - 0.92	F/C: high ; Co: ICC : FSQ: *r* : 0.46 Oswestry: :*r* : 0.38 PPM: Obstacle course : *r* : 0.48 – 0.78 Re : sign. ↓ pre- and post- operative	Salén B.A. et al 1994 [54]
Medical Rehabilitation Follow Along (MRFA) [55]			I.C.C.: 0.74 – 0.97	κ: 0.52 – 0.66		Granger C.V. et al 1995 [55]
MOS 36-item Short Form Health survey (MOS 36-SF) [104]				*r*: 0.43-0.90	Co: high Cr: MOS 36 – QBS: *r*: 0.72	McHorney C et al 1993 [104] Harwood K.J. 2001 [67]
Million Visual Scale [57]	*r*: 0.92	*r*: 0.97			Co: MVAS- VAS pain: r: 0.44 - pain/impairment scale: r: 0.79	Million R. et al 1982 [57] Anagnostis C. et al 2003 [105]

					Re: OR 1.7 pretreatment OR 3.1 posttreatment	Beurskens A.J. et al 1995 [83]
Oswestry Disability Questionnaire/ Index (ODI) [24]			α: 0.71-0.87	*r*: 0.99 1-day *r*: 0.91 4-days *r*: 0.83 7-days	F/C: Change: improvement sign. ↓ Co: ODI- VAS: *r* : 0.64 RDQ: *r*: 0.77 PDI: *r*: 0.83 QBS: *r*: 0.80 Re: ROC index: 0.76	Fairbank J.C.T. et al 1980 [24] Roland M. et al 2000 [82] Beurskens A.J. et al 1995 [83]
Questionnaire Physical Activities [59]			ICC: 0.49-0.94			Torgen M. et al 1997 [59]
Pain Disability Index (PDI) [58]			α: 0.86	PPM: *r*: 0.44 2-month	Cr: Pain ↑ high PDI group Co: PDI- ODI: *r*: 0.83 RMDQ: *r*: 0.59/0.63 SFS: *r*: -.64 VAS: *r*: .54 Multiple regression: Multi. *R:* 0.74 (54% of total)	Tait R.C. et al 1990 [58] Beurskens A.J. et al 1995 [83]
Roland Morris Disability Questionnaire (RMDQ) [25]			α: 0.84-0.93	*r*: 0.91 1 day *r*: 0.88 7 days *r*: 0.83 21 days	F/C: moderate Co: RMDQ- VAS pain: *r*: 0.47/0.62 - SIP: *r*: 0.78-0.89 - Quebec Back Scale: *r*: 0.77 - Oswestry: *r*: 0.77 - PDI: *r*: 0.59/0.63 Re SRM: 0.77 Area under the ROC: 0.73	Roland M. et. al 1983 [25] Beurskens A.J. et al 1995 [83] Stucki G. et al 2000 [85] Jensen M.P. et al 1992 [113]
Spinal Function Sort (SFS) [60]			ICC: 0.89 α: .98		Co: SFS- other scales; *r*: -.64 -.78 sign. Multiple regression: Multi *R*: .63 (72% of total)	Gibson L. et al 1996 [60]
Neck Disability Index [61]			α: 0.80	*r*: 0.89	Cr : NDI- VAS: *r*: 0.60 NDI- MPQ: *r*: 0.70 Co: normal distribution: 83% mild-moderate categories	Vernon H. et al 1991 [61] Ackelman B.H.& Lingren U. 2002 [114]
Upper Extremity Function Scale (UEFS) [63]			α: 0.83-0.93		Cr: UEFS-AIMS: *r*: 0.81 UEFS: UED-CTS: differences sign. + Re: corr.: longitudinal measures – UEFS : sign. +	Pransky G. et al 1997 [63]
Lower Extremity Activity Profile (LEAP) [64]			α: 0.73		Co: corr. LEAP-SPW: low- moderate corr. LEAP-ROM: not sign. Re: change: pre- post operative: sign. ↑	Finch E. et al 1995 [64]
Functional tests						
Baltimore Therapeutic Equipment [65]			*r*: 0.62- 0.82			Bhambhani Y. et al 1993 [65]
DOT Residual Functional Capacity [110]	-	-	-	-	Cr: % corr.class.: 61.1-79.4 % sensitivity 69.5- 100 % specificity 27.3- 74.6	Fishbain D.A. et al. 1999 [110]
Functional Capacity Evaluation [68]	-	-	-	-	Co *r:* - 0.4821 standing: sign. Other: not sign.	Parks K.A. et al 2003 [68]
Physical Functional Test [69]	-	-	-	-	Co: *r*: PFS: - Oswestry: 0.197 – FCE: -.154 – 0.051	Hart D.L. 1998 [69]
Physical Performance Test (BPS : Back Performance Scale) [80]			α: .73		Cr: higher BPS : sign. + : more pain Bivariate corr BPS: *r*: .63 - . 73 BPS-tests: *r*: .63- .73 Re: sensitivity: 67%; specificity: 70%. Cutoff point : 2.5	Strand L.I. et al 2002 [80]
Tufts Assessment of Motor Performance (TAMP) [71]			ICC: 0.71-0.99 κ: 0.63- 0.84			Gans B.M. et al 1988 [71]
EPIC Lift Capacity test [72,111]		PPM: *r*: .90	ICC: ..91		Re Reactivity: Before-after treatment: sign ↑	Matheson L.N. et al. 1995 [72,111]
Lifting tests [73]	-	-	-	-	Co: Leg lift: sign ↓ norm Arm lift: not sign Trunk lift: not sign.	Leferink V.J.M. et al 2003 [73]
Physical Work Capacity [74]	-	-	-	-	Cr: corr. all PWC variables: 0.81 –0.97 corr. Borg rating- lift weight: sign. +	Jackson A.S. et al 1997 [74]
Progressive Isoinertial Lifting Evaluation PILE [29]	ICC: lumbar: 1.0 cervical: 1.0	ICC: lumbar: 0.70 cervical: 0.92			Cr: Lumbar: 2 groups sign. ↑ lifting weights	Horney E. et al 2002 [29] Mayer T. et al 1994 [75]
Jebsen Hand Function Test [28]		*r*: 0.99		P.P.M. *r*: .60- ..99	Cr: Free-immobilised hand: sign. ↓ less time	Jebsen R.H. et al 1969 [28] Chan W.Y.Y. & Chapparo C. 1999 [112]
Functional Performance Tests [77]				ICC: .66- ..96	Cr: Injured-unjured limb: difference not sign.	Bolgla L.A. et al 1997 [77] Munn J. et al. 2002 [78]
Motor Activity Score [79]			ICC .90		Re: Athlability + Activity: difference: sign. ↑ postinjury days	Wilson R.W. et al 1998 [79]

FSQ :Functional Status Questionnaire ; VAS:Visual Analogue Scale ; QBS :Quebec Back Scale;SRM :Standardized Response Mean; SIP:Sickness Impact Profile; MPQ:McGill Pain Questionnaire; AIMS:Arthritis Impact Measurement Scale;UED:Upper Extremity Disorder; CTS:Carpal Tunnel Syndrome;SPW:Self Paced Walk;ROM:Range of Motion

2.4 Discussion

The purpose of the present review was to present an overview of methods to assess the functional capacity of the musculoskeletal system. In order to obtain the available literature on this topic we systematically searched the literature in eight databases. A total of 48 original studies were included. In these studies, 13 questionnaires and 14 functional tests were described. Of the questionnaires, the Pain Disability Index [58], the Oswestry Disability Questionnaire [24], the Roland Morris Disability Questionnaire [25] , and the UEFS [63] had high levels of both reliability and validity. Of the functional tests, none had high levels of both reliability and validity. Ten out of 13 questionnaires were used in the context of work. Three questionnaires focussed especially on patients with low back pain, but one focussed on patients with disorders of the upper extremities [63].

As far as we know no previous study was performed to present an inventory of possible methods to assess the functional capacity of the musculoskeletal system. Despite the systematic nature of this review and the great number of databases used, some relevant studies may not have been included. However, because the references of the studies included were also used, we presume that this number is limited. Nevertheless, we are aware that a number of questionnaires and functional tests are employed to assess functional capacity of subjects with musculoskeletal disorders that were not published in peer-reviewed journals.

The inclusion criteria consisted of four criteria. Two will be addressed shortly: the context and the functional assessment method. The context is important because the context determines whether the reported impairment leads to restrictions and limitations in participation and activities in accordance with the ICF model [2]. In 423 studies no context was specified and 393 studies failed to describe a functional assessment method. Therefore, a great number of studies were excluded. Many studies were excluded because the assessment methods were only directed at finding or confirming a diagnosis. Besides, many assessment methods were only used to evaluate the results of therapy in terms of exerted force or range of motion. In these studies neither context nor a functional assessment method was described.

We chose the rating system of Hulshof et al [33] for appraisal of the methodological quality of the studies. According to the levels of quality rating, a large number of studies were qualified as moderate or poor. Without reliability and validity, the quality of an assessment method is at least questionable. Therefore, these studies were not discussed. A meta analysis could not be performed, because there was not enough homogeneity in the studies, which is a prerequisite for a meta analysis.

Practical relevance

What methods should be used in practice to assess the functional capacity of the musculoskeletal system? The present study shows that three questionnaires have a high level of reliability and validity. No reliable and valid functional tests were found. The questionnaires contain mainly questions about activities of daily living. Though activities of daily living and work are overlapping, the translation of scores from these three questionnaires to functional capacity for work could be doubted. In many work situations more physically demanding activities than in daily life have to be performed, in terms of not only level but also frequency and duration. The ability of subjects with musculoskeletal disorders cannot be assessed on the basis of the questionnaires alone. The questionnaires often lack information on the level and duration of these activities. Some activities, such as kneeling, reaching, and pushing and pulling, are not or not extensively rated in the questionnaires, whereas they are essential activities in many jobs. Several authors describe the ability of a functional test to assess the functional capacity of workers [75,86-88]. A functional test may provide clarity about, for instance, level, frequency and duration of activities and fills in the lacking exposure information of the questionnaires. The results of these functional tests [87,89] are influenced by conditions, such as fluctuation of performance during the day and between days and, variable course of some medical condition. Besides, there may be ambiguity about the level of performance and the sincerity of effort [87,90,91].

Important in the context is the influence of pain, fear of pain, fear of re-injury, but also depression, anxiety, somatization and other major psychosocial barriers, related to the ability to perform work-related tasks [92,93]. Self-efficacy is proven to be of great influence towards actual functioning. The goals that are set for task performances, along with performance self-efficacy expectancies, have a direct and independent influence on task performance [94].

Then also, the purpose of the assessment such as an evaluation of rehabilitation or an insurance claim might influence the outcome of the assessment. Therefore, a combination of different methods of measurement seems to be the most desirable in order to achieve a correct assessment, though this was not tested empirically. The outcome of the different assessments may be combined, leading to a consistent and complete judgment. This should be further investigated. Until now, a reliable and valid set of tools for the purpose of evaluation of human function related to musculoskeletal pain and impairment is still missing [95].

The questionnaires that were selected apply to populations of patients with general disorders. As a consequence, for groups of patients with specific disorders, such as malfunction of the hand, and knee or ankle injuries, these general questionnaires could be of

limited use. Perhaps, it is appropriate to choose a more specific questionnaire in case of a specific disorder [28,29,63].

Finally, for the assessment of the functional capacity of low back patients a number of reliable and valid questionnaires are available [96]. These questionnaires are pre-eminently useful in the context of work, but also seem useful in the context of daily activities. For assessment of upper extremity disorders in the context of work, the UEFS [63] can be used as a reliable and valid questionnaire, useful in the context of work. For sport, only functional tests were found that were reliable but insufficiently validated. When we focus on work, we need a set of tests that assess the general functional capacity of the musculoskeletal system that have a sufficient validity and that can be used in combination with the selected questionnaires.

Reference list

1. Albright A, Franz M, Hornsby G, Kriska A, Marrero D, Ullrich I, Verity LS (2000) American College of Sports Medicine position stand. Exercise and type 2 diabetes. Med Sci Sports Exerc 32(7): 1345-1360
2. WHO (2001) International Classification of Functioning, Disability and Health: ICF; Geneva
3. Van Tulder M, Koes BW, Bouter LM (1995) A cost of illness study of back pain in the Netherlands. Pain 62: 233-240
4. Borghouts JAJ, Koes BW, Vondeling H, Bouter LM (1996) Cost-of-illness of neck pain in the Netherlands in 1996. Pain 80: 629-636
5. Hemmilä HM (2002) Quality of life and cost of care of back pain patients in Finnish general practice. Spine 27: 647-653
6. Picavet HSJ , Schouten JSAG (2003) Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC3-study. Pain 102: 167-178
7. Urwin M, Symmons D, Allison T, Brammah T, Busby H, Roxby M, Simmons A, Gareth G (1998) Estimating the burden of musculoskeletal disorders in the community: the comparative prevalence of symptoms at different anatomical sites, and the relation to social deprivation. Ann Rheum Dis 57: 649-655
8. Reginster JY (2002) The prevalence and burden of arthritis. Rheumatology 41 (suppl 1): 3-6
9. Picavet HSJ, Schouten JSAG (2000) Physical load in daily life and low back problems in the general population- The MORGEN study. Preventive Medicine 31: 506-512
10. Ariens GA, Bongers PM, Douwes M, Miedema MC, Hoogendoorn WE, van der Wal G, Bouter LM, van Mechelen W (2001) Are neck flexion, neck rotation, and sitting at work risk factors for neck pain? Occup Environ Med 58: 200-207
11. Burdof A , Sorock G (1997) Positive and negative evidence of risk factors for back disorders. Scand J Work Environ Health 23: 243-256
12. Hoogendoorn WE, Bongers PM, de Vet HC, Douwes M, Koes BW, Miedema MC, Ariens GA, Bouter LM (2000) Flexion and rotation of the trunk and lifting at work are risk factors for low back pain: results of a prospective cohort study. Spine 25(23): 3087-3092
13. Hoozemans MJM, van der Beek AJ, Frings-Dresen MHW, van der Woude LHV, van Dijk FJH (2002) Pushing and pulling in association with low back and shoulder complaints. Occup Environ Med 59: 696-702

14. Feuerstein M (1990) A multidisciplinary approach to the prevention, evaluation, and management of work disability. J Occup Rehabil 1 (1): 5-12
15. Lundgren-Lindquist B, Sperling L (1983) Functional studies in 79-year-olds. II. Upper extremity function. Scand J Rehabil Med Vol 15(3): 117-123
16. Wildner M, Wildner M, Sangha O, Clark DE, Doring A, Manstetten A (2002) Independent living after fractures in the elderly. Osteoporos Int 13(7): 579-585
17. Van Schaardenburg D, van den Brande KJ, Ligthart GJ, Breedveld FC, Hazes JM (1994) Musculoskeletal disorders and disability in persons aged 85 and over: a community survey. Ann Rheum Dis 53(12): 807-811
18. Hootman JM, Macera CA, Ainsworth BE, Addy CL, Martin M (2002) Epidemiology of musculoskeletal injuries among sedentary and physically active adults. Med Sci Sports Exerc 34(5): 838-844
19. Marshall SW, Mueller FO, Kirby DP, Yang J (2003) Evaluation of safety balls and faceguards for prevention of injuries in youth baseball. JAMA 289 (Feb 5): 194-195
20. Abernethy L, MacAuley D (2003) Impact of school sports injury. Br J Sports Med 37: 354-355
21. Gabbett TJ (2003) Incidence of injury in semi-professional rugby league players. Br J Sports Med 37: 36-43
22. Federiuk CS, Schlueter JL, Adams AL (2002) Skiing, snowboarding, and sledding injuries in a northwestern state. Wilderness Environ Med 13: 245-249
23. Boyce SH, Quigley MA (2003) An audit of sports injuries in children attending an Accident & Emergency department. Scott Med J 48: 88-90
24. Fairbank JCT, Couper J, Davies JB, O'Brien JP (1980) The Oswestry low back pain questionnaire. Physiotherapy 66: 271-273
25. Roland M, Morris R (1983) A study of natural history of low back pain. Part 1: Development of a reliable and sensitive measure of disability in low-back pain. Spine 8: 141-144
26. Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, Williams JL (1995) The Quebec back pain disability scale. Spine 20: 341-352
27. Tramposh AK (1992) The functional capacity evaluation: measuring maximal work abilities. Occup Med 7(1): 113-124
28. Jebsen RH, Taylor N, Trieschmann RB, Trotter MJ, Howard LA (1969) An objective and standardized test of hand function. Arch Phys Med Rehabil 50(6): 314-319

29. Horneij E, Holmström E, Hemborg B, Isberg P-E, Ekdahl C (2002) Interrater reliability and between days repeatability of eight physical performance tests. Adv Phys 4: 146-160
30. Millard RW (1991) A critical review of questionnaires for assessing pain-related disability. J Occup Rehabil 1(4): 289-302
31. Altman DG (1991) The medical literature. In Chapman & Hall, editor. Practical statistics for Medical Research. London, New York, Tokyo, Melbourne, Madras pp 477-499
32. Hoogendoorn WE, van Poppel MNM, Bongers PM, Koes BW, Bouter LM (1999) Physical load during work and leisure time as risk factors for back pain: a systematic review. Scand J Work Environ Health 25: 387-403
33. Hulshof CTJ, Verbeek JHAM, van Dijk FJH, van der Weide WE, Braam ITJ (1999) Evaluation research in occupational health services: general principles and a systematic review of empirical studies. Occup Environ Med 56: 361-377
34. Streiner DL, Norman GR (2003) Chapter 8 Reliability. Health Measurement Scales: A practical guide to their development and use. Oxford 3th: p 126
35. Carmines EG, Zeller RA. Reliability and validity assessment. Newbury Park, London, New Dehli; Sage publications 1979 pp 11-16
36. Innes E, Straker L (1999) Reliability of workrelated assessments. Work 13: 107-124
37. Carmines EG, Zeller RA (1979) Reliability and validity assessment. Newbury Park, London, New Dehli. Sage publications pp 37-51
38. Carmines EG, Zeller RA (1979) Reliability and validity assessment. Newbury Park, London, New Dehli. Sage publications pp 17-27
39. Bouter LM, van Dongen MCIM (2000) Epidemiological research; design and interpretation [Epidemiologisch onderzoek; opzet en interpretatie: in Dutch]. Houten/Diegem: Bohn Stafleu van Loghum. Chapter 4 pp 279-279
40. Streiner DL, Norman GR (2003) Health measurement scales. A practical guide to their development and use. Chapter 7: From items to scales; Oxford 3th pp 119-122
41. Airaksinen O, Herno A, Saari T (1994) Surgical treatment of lumbar spinal stenosis: patients' postoperative disability and working capacity. Eur Spine J 3(5): 261-264
42. Burd TA, Pawelek L, Lenke LG (2002) Upper extremity functional assessment after anterior spinal fusion via thoracotomy for adolescent idiopathic scoliosis: prospective study of twenty-five patients. Spine 27(1): 65-71
43. Hodges SD, Humphreys SC, Eck JC, Covington LA, Harrom H (2001) Predicting factors of successful recovery from lumbar spine surgery among workers' compensation patients. J Am Osteopath Assoc 101(2): 78-83

44. Lyle RC (1981) A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res 4(4): 483-492
45. Milhous RL, Haugh LD, Frymoyer JW, Ruess JM, Gallagher RM, Wilder DG, Callas PW (1989) Determinants of vocational disability in patients with low back pain. Arch Phys Med Rehabil 70(8): 589-593
46. Pennathur A, Mital A, Contreras LR (2001) Performance reduction in finger amputees when reaching and operating common control devices: a pilot experimental investigation using a simulated finger disability. J Occup Rehabil 11(4): 281-290
47. Weiss AC, Wiedeman G, Quenzer D, Hanington KR, Hastings H, Strickland JW (1995) Upper extremity function after wrist arthrodesis. J Hand Surg [Am] 20(5): 813-817
48. Wolf LD, Matheson LN, Ford DD, Kwak AL (1996) Relationships among grip strength, work capacity, and recovery. J Occup Rehabil 6(1): 57-70
49. Mayer TG, Gatchel RJ, Kishino N, Keeley J, Capra P, Mayer H, Barnett J, Mooney V (1985) Objective assessment of spine function following industrial injury. A prospective study with comparison group and one-year follow-up. Spine 10(6): 482-493
50. Rayan GM, Brentlinger A, Purnell D, Garcia-Moral CA (1987) Functional assessment of bilateral wrist arthrodeses. J Hand Surg [Am] 12(6): 1020-1024
51. Saunders RL, Beissner KL, McManis BG (1997) Estimates of weight that subjects can lift frequently in functional capacity evaluations. Phys Ther 77(12): 1717-1728
52. Irrgang JJ, Snyder-Mackler L, Wainner RS, Fu FH, Harner CD (1998) Development of a patient-reported measure of function of the knee. J Bone Joint Surg 80: 1132-1145
53. Gronblad M, Jarvinen E, Hurri H, Hupli M, Karaharju EO (1994) Relationship of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ)with three dynamic physical tests in a group of patients with chronic low-back and leg pain. Clin J Pain 10(3): 197-203
54. Salen BA, Spangfort EV, Nygren AL, Nordemar R (1994) The Disability Rating Index: an instrument for the assessment of disability in clinical settings. J Clin Epidiomol 47(12): 1423-1435
55. Granger CV, Ottenbacher KJ, Baker JG, Sehgal A (1995) Reliability of a brief outpatient functional outcome assessment measure. Am J Phys Med Rehabil 74(6): 469-475
56. Ware JE, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). Med Care 30(6): 473-481
57. Million R, Hall W, Nilsen KH, Baker RD, Jayson MIV (1982) Assessment of the progress of the back pain patient. Spine 7: 204-212

58. Tait RC, Chibnall JT, Krause S (1990) The pain disability index: psychometric properties. Pain 40: 171-182
59. Torgen M, Alfredsson L, Köster M, Wiktorin C, Smith KF, Kilbom A (1997) Reproducibility of a questionnaire for assessment of present and past physical activities. Int Arch Occup Environ Health 70: 107-118
60. Gibson L, Strong J (1996) The reliability and validity of a measure of perceived functional capacity for work in chronic back pain. J Occup Rehabil 6(3): 159-175
61. Vernon H , Mior S (1991) The neck disability index: a study of reliability and validity. J Manip Physiol Ther 14(7): 409-415
62 .Carroll D (1965) A quantitative test of upper extremity function. J Chron Dis 18: 479-491
63. Pransky G, Feuerstein M, Himmelstein J, Katz JN, Vickers LM (1997) Measuring functional outcomes in work-related upper extremity disorders - Development and validation of the upper extremity function scale. J Occup Environ Med 39: 1195-1202
64. Finch E, Kennedy D (1995) The lower extremity activity profile: a health status instrument for measuring lower extremity disability. Physiother Can 47(4): 239-246
65. Bhambhani Y, Esmail S, Brintnell S (1993) The Baltimore Equipment Work Simulator: Biomechanical and physiological norms for three attachments in healthy men. Am J Occup Med 48(1): 19-25
66. Fishbain DA, Abdel ME, Cutler R, Khalil TM, Sadek S, Rosomoff RS, Rosomoff HL (1994) Measuring Residual Functional Capacity in Chronic Low Back Pain Patients Based on the Dictionary of Occupational Titles. Spine 19: 872-880
67. Harwood KJ (2001) The process of returning to work following an episode of disabling low back pain: a phenomenological study. New York University ** Ph D.(102 p)
68. Parks KA, Crichton KS, Goldford RJ, McGill SM (2003) A comparison of lumbar range of motion and functional ability scores in patients with low back pain: assessment for range of motion validity. Spine 28(4): 380-384
69. Hart DL (1998) Relation between three measures of function in patients with chronic work-related pain syndromes. J Rehabil Outcome Meas 2(1): 1-14
70. Strand LI, Ljunggren AE (2001) The pick-up test for assessing performance of a daily activity in patients with back pain. Adv Physiother 3(1): 17-27
71. Gans BM, Haley SM, Hallenborg SC, Mann N, Inacio CA, Faas RM (1988) Description and interobserver reliability of the Tufts Assessment of Motor Performance. Am J Phys Med Rehabil 67(5): 202-210

72. Matheson LN, Mooney V, Grant JE, Affleck M, Hall H, Melles T, Lichter RL, McIntosh G (1995) A test to measure lift capacity of physically impaired adults. Part 1: development and reliability testing. Spine 20(19): 2119-2129
73. Leferink VJM, Keizer HJE, Oosterhuis JK, Van der Sluis CK, Ten Duis HJ (2003) Functional outcome in patients with thoracolumbar burst fractures treated with dorsal instrumentation and transpedicular cancellous bone grafting. Eur Spine J 12(3): 261-267
74. Jackson AS, Borg G, Zhang JJ, Laughery KR, Chen J (1997) Role of physical work capacity and load weight on psychophysical lift ratings. Int J Ind Erg 1997 20(3): 181-190
75. Mayer T, Gatchel R, Keeley J, Mayer H, Richling D (1994) A male incumbent worker industrial database. Part III: Lumbar/ cervical functional testing. Spine 19(7): 765-770
76. Barber SD, Noyes FR, Mangine RE (1991) Quantitative assessment of functional limitations in normal and anterior cruciate ligament-deficient knees. Clin Orthop Rel Res 255: 204-241
77. Bolgla LA, Keskula DR (1997) Reliability of lower extremity functional performance tests. J Orthop Sports Phys Ther 26(3): 138-142
78. Munn J, Beard DJ, Refshauge KM, Lee RWY (2002) Do functional-performance tests detect impairment in subjects with ankle instability? J Sport Rehabil 11(1): 40-50
79. Wilson RW, Gieck JH, Gansneder BM, Perrin DH, Saliba EN, McCue FC (1998) Reliability and responsiveness of disablement measures following acute ankle sprains among athletes. J Orthop Sport Phys 27(5): 348-355
80. Strand LI, Moe-Nilssen R, Ljunggren AE (2002) Back Performance Scale for the assessment of mobility-related activities in people with back pain. Phys Ther 82(12): 1213-1223
81. Stratford PW, Binkley JM, Riddle DL, Guyatt GH (1998) Sensitivity to change of the Roland-Morris back pain questionnaire: part 1. Phys Ther 78(11): 1186-1196
82. Roland M, Fairbank J (2000) The Roland-Morris disability questionnaire and the Oswestry disability questionnaire. Spine 25(24): 3115-3124
83. Beurskens AJ, de Vet HC, Köke AJ, van der Heijden AG, Knipschild PG (1995) Measuring the functional status of patients with low back pain. Spine 20(9): 1017-1028
84. Riddle DL, Stratford PW, Binkley JM (1998) Sensitivity to change of the Roland-Morris pain questionnaire: part 2. Phys Ther 78(11): 1197-1207
85. Stucki G, Kroeling P (2000) Physical therapy and rehabilitation in the management of rheumatic disorders. Best Pract Res Clin Rheumatol 14(4): 751-771

86. Wyman DO (1999) Evaluating patients for return to work. Am Fam Physician 59(4): 844-848
87. Strong S (2002) Developing expert practice. Functional capacity evaluation: the good, the bad and the ugly. Occup Ther Now 4: 5-9
88. Lechner DE, Jackson JR, Roth DL, Straaton KV (1994) Reliability and validity of a newly developed test of physical work performance. J Occup Med 36: 997-1004
89. Wijnen JAG, Boersma MThLW (2001) Disability claim assessement and assessing the functional capacity [in Dutch: Claimbeoordeling en het bepalen van de functionele capaciteit] Tijdschr Bedrijfs Verzekeringsgeneeskd (3): 70-71
90. Lechner DE, Bradbury SF, Bradley LA (1998) Detecting sincerity of effort: a summary of methods and approaches. Phys Ther 78: 867-888
91. Simonsen JC (1996) Validation of sincerity of effort. J Back Musculoskelet 6: 289-295
92. Papciak AS, Feuerstein M (1991) Psychological factors affecting isokinetic trunk strength testing in patients with work-related chronic low back pain. J Occup Rehabil 1(2): 95-104
93. Gatchel RJ (2004) Psychosocial factors that can influence the self-assessment of function. J Occup Rehabil 14(3): 197-206
94. Lackner JM, Carosella AM, Feuerstein M (1996) Pain expectancies, pain, and functional self-efficacy expectancies as determinants of disability in patients with chronic low back disorders. J Consul Clin Psychol 64(1): 212-220
95. Feuerstein M (2004)Functional assessment for persons with musculoskeletal pain and impairment. J Occup Rehabil 14(3): 163-164
96. Deyo RA, Battie M, Beurskens AJHM, Bombardier C, Croft P, Koes B, Malmivaara A, Roland M, Von Korff M, Waddell G (1998) Outcome measures for low back pain research. Spine 23(18): 2003-2013
97. Nunnally JC (1978) Psychometric theory. New York: McGraw-Hill 3rd
98. Altman DG (1991) Some common problems in medical research. In Chapman&Hall, editor. Practical statistics for medical research. London, New York, Tokyo, Melbourne, Madras; 1th(14): 404.
99. Innes E, Straker L (1999) Validity of work-related assessments. Work 13: 125-152
100. Van den Hout WB (2003) The area under an ROC Curve with Limited Information. Med Decis Making 23: 160-166
101. Obuchowski NA (2003) Receiver operating characteristic curves and their use in radiology. Radiology 229: 3-8

102.Deyo RA, Centor RM (1986) Assessing the responsiveness of functional scale to clinical change: an analogy to diagnostic test performance. J Chron Dis 39(11): 897-906

103.Portney LG, Watkins MP (2000) Foundations of clinical research: Application to practice. Norwalk, Connecticut; Appleton&Lange

104.Mc Horney C,Ware JE, Raczek AE (1993) The MOS 36-item short-form survey (SF-36: II Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 31: 247-263

105.Anagnostis C, Mayer TG, Gatchel RJ et al (2003). The Million Visual Analog Scale: Its utility for predicting tertiary rehabilitation outcomes. Spine 28: 1051-1060

106.Di Fabio RP, Mackey G, Holte JB (1996) Physical therapy outcomes for patients receiving worker's compensation following treatment for herniated lumbar disc and mechanical low back pain syndrome. J Orthop Sports Phys Ther 23: 180-187

107.Loisel P, Poitras S, Lemaire J, Durand P, Southiere A, Abenhaim L (1998) Is work status of low back pain patients best described by an automatic device or by a questionnaire? Spine 23: 588-594

108.Poitras S, Loisel P, Prince F, Lemaire J (2000) Disability measurement in persons with back pain: a validity study of spinal range of motion and velocity. Arch Phys Med Rehabil 81(10): 1394-1400

109.Torgen M, Punnett L, Alfredsson L, Kilbom (1999) Physical capacity in relation to present and past physical load at work: a study of 484 men and women aged 41 to 58 years. Am J Ind Med 36: 388-400

110.Fishbain DA, Cutler RB, Rosomoff H, Khalil T, Abdel-Moty E, Steele-Rosomoff R (1999) Validity of the dictionary of occupational titles residual functional capacity battery. Clin J Pain15: 102-110

111.Matheson LN, Mooney V, Holmes D, Leggett S, Grant JE, Negri S (1995) A test to measure lift capacity of physically impaired adults. Part 2: reactivity in a patient sample. Spine 20: 2130-2134

112.Chan WYY, Chapparo C (1999) Effect of wrist immobilisation on upper limb function of elderly males. Technol Disabil 11: 39-49

113.Jensen MP, Strom SE, Turner JA, Romano JM (1992) Validity of the sickness impact profile roland scale as a measurement of dysfunction in chronic pain patients. Pain 50: 157-162

114.Ackelman BH, Lindgren U (2002) Validity and reliability of a modified version of the neck disability index. J Rehabil Med 34: 284-2

Chapter 3

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

Vincent Gouttebarge, Haije Wind, P.Paul.F.M. Kuijer, Judith K. Sluiter, Monique H.W. Frings-Dresen; International Archives of Occupational and Environmental Health 2004; 77: 527-537

Abstract

Objectives

Functional Capacity Evaluation methods (FCE) claim to measure the functional physical ability of a person to perform work-related tasks. The purpose of the present study was to systematically review the literature on the reliability and validity of four FCEs: the Blankenship System (BS), the ERGOS Work Simulator (EWS), the Ergo-Kit (EK) and the Isernhagen Work System (IWS).

Methods

A systematic literature search was conducted in five databases (CINAHL,Medline, Embase, OSH-ROM and Picarta) using the following keywords and their synonyms: functional capacity evaluation, reliability and validity. The search strategy was performed for relevance in titles and abstracts, and the databases were limited to literature published between 1980 and April 2004. Two independent reviewers applied the inclusion criteria to select all relevant articles and evaluated the methodological quality of all included articles.

Results

The search resulted in 77 potential relevant references but only 12 papers were identified for inclusion and assessed for their methodological quality. The interrater reliability and predictive validity of the IWS were evaluated as good while the procedure used in the intrarater reliability (test–retest) studies was not rigorous enough to allow any conclusion. The concurrent validity of the EWS and EK was not demonstrated while no study was found on their reliability. No study was found on the reliability and validity of the BS.

Conclusions

More rigorous studies are needed to demonstrate the reliability and the validity of FCE methods, especially the BS, EWS and EK.

3.1 Introduction

In a world that is changing continuously and where everything is moving faster, functioning as a human being is really important. All human movement, from laughing to walking, depends on the proper functioning of our musculoskeletal system. This complex system allows us to perform different tasks in daily life, for instance at work. The musculoskeletal system has been identified as the most common cause of occupational disease and work loss: it especially concerns disorders such as low back pain, neck pain, upper limb pain and arthritis [1-4]. In recent years, as the incidence of work-related injuries and occupational diseases has risen considerably, there has been growing interest in musculoskeletal disorders in workers. Reducing work-related injuries or illness, and their medical costs, has become a priority in many countries.

In the Netherlands, work disability, defined as the inability to perform job tasks as a consequence of physical or mental unfitness, became over the last decades a socio-economic problem and actually dominates the political debate. From 1976 to 2001, the number of injured or sick workers who were partially or fully disabled for work and received work compensation rose for more than 50%, growing to almost 1 million people, and that for a substantial work population of 8.5 million people [5,6]. The total healthcare cost for this large number of people with work disability reaches each month 850 million euros, representing an expenditure of more than 10 milliard of euros over a whole year [6]. Impairments of the musculoskeletal system are, beside the psychological disorders, the most important causes responsible for disability and work absenteeism: 36% of all people seen during a work disability claim for work compensation had an occupational disorder or injury related to the musculoskeletal system [6].

Functional Capacity Evaluation (FCE) aims to be a systematic, comprehensive and multi-faceted ‘‘objective’’ measurement tool designed to measure someone’s current physical abilities in work-related tasks [7-9]. FCEs are commonly used for individuals who have work-related disorders, particularly musculoskeletal disorders [9,10]. FCEs are used by physicians, insurance companies, medical care organizations as well as in industry and government entities during work disability claims, injury prevention, rehabilitation process, work conditioning programs, return to work decision after injury and pre-employment screening for people with or without impairments [11-12]. Over the past few years, a number of FCEs has been developed to assess functional capacity in specific work-related tasks. In the Netherlands, four

major FCEs are developing and profiling themselves on the Dutch market as high quality work assessment methods: Blankenship System (BS)[13], Ergos Work Simulator (EWS)[14], Ergo-Kit (EK)[15] and Isernhagen Work System (IWS)[16].

For these four FCEs, the principles of scientific measurement should be considered, as they are for any other test: an FCE should give reliable and valid measurements[17]. The providers of these FCEs pretend that these assessments use procedures that are reliable and valid[18]. However, they do not supply enough evident information about the reliability and validity of these FCEs. Gardener et al. even notices that the lack of documented reliability and validity diminishes confidence in any approach to FCE[19].

The aim of the present study is to review systematically the literature on the reliability and validity of the BS, EWS, EK and IWS. This objective results in the following questions:

(a) What is known about the reliability of the BS, EWS, EK and IWS?

(b) What is known about the validity of the BS, EWS, EK and IWS?

3.2 Methods

Systematic search strategy

We performed a systematic literature search involving the following electronic databases: CINAHL (nursing and allied health literature), Medline (biomedical literature), Embase (biomedical and pharmacological literature) and OSH-ROM (occupational safety and health related literature, including databases as RILOSH, MIHDAS, HSELINE, CISDOC and NIOSHTIC2).

We used the following keywords and their synonyms: functional capacity evaluation combined with reliability / validity (Table 1). The synonyms of functional capacity evaluation were connected by ‘‘or’’, so as the synonyms for reliability and validity. Both groups of results were then connected by ‘‘and’’.

The search strategy was performed for relevance in titles and abstracts, and the databases were limited to literature published between 1980 and April 2004. We also searched a Dutch database, Picarta, to identify publications written in Dutch using as keywords the names of the four FCEs: Blankenship, Ergos, Ergo-Kit, and Isernhagen.

Table 1: Key words and their synonyms used in the present study

Functional Capacity Evaluation	Reliability / Validity
Functional capacity evaluation	Reliability
FCE	Reliable
Blankenship	Repeatable
Ergos	Reproducibility
Ergo-kit	Test-retest
Isernhagen	Intrarater reliability
	Interrater reliability
	Consistency
	Consistent
	Stability
	Precision
	Validity
	Valid

Inclusion criteria

Inclusion criteria were defined and used to ensure capturing all relevant literature. We included articles:

(1) written in English, Dutch or French

(2) and using one of the following FCE's: Blankenship, Ergos, Ergo-Kit, Isernhagen

(3) and presenting data about the reliability and/or validity of these FCE's.

Study selection

Applying the inclusion criteria defined above, the first two authors independently reviewed the titles and abstracts of the literature to identify potentially relevant articles (step 1). If any title and abstract did not provide enough information to decide whether or not the inclusion criteria were met, the article was included for the full text selection. From the titles and abstracts included, we read the full articles and the same two reviewers applied the inclusion criteria to the full text (step 2). Disagreements, if any, on the inclusion or exclusion of articles were resolved by consulting a third reviewer.

Reviews were included and only used to screen for further original papers. The bibliographies of the articles included were also cross-checked to search for studies not referenced in our databases as we systematically searched for the name of one of the four FCEs (Blankenship, Ergos, Ergo-Kit, Isernhagen) in the titles of the references. Then, we applied the three inclusion criteria to the full text.

Methodological quality appraisal

All included articles were reviewed independently by the first two authors to assess the methodological quality. As the methodological quality in a study influences the results and conclusions in our systematic review, we developed a three-level quality appraisal scale (+,± and -) to evaluate the scientific relevance of each study. This scale is, for a large part, based on different studies [20-25].

Five methodological quality appraisal features were defined and assessed: (1) *functional capacity evaluation* to evaluate if it is clearly mentioned whether the full FCE method has been used or which subtests, (2) *objective* to evaluate whether the objective of the study is clearly defined, (3) *study population* to judge whether the study population is well described, (4) *procedure* to evaluate whether the study used a properly defined procedure to achieve the objective [21-25], and (5) *statistics* to evaluate whether the statistics used are clearly described and properly used to test the hypothesis of the study [20].

Each study get 5 scores and the total score was calculated by adding + and – scores: +, +, ±, +, - give a total of 2 +, as one – eliminates one + and +/- does not count. The methodological quality of the studies is rated as follow:

- high: 4 or 5 +, indicating a high methodological quality,
- moderate: 2 or 3 +, indicating a moderate methodological quality,
- and low: 0 or 1 +, indicating a low methodological quality.

Any disagreement between both reviewers was resolved by consulting a third reviewer. Table 2 gives a completed description of these methodological quality appraisals.

Table 2: The methodological quality appraisal [21-25]

1.	FCE method	
	+	It is clearly mentioned in this study whether the full FCE-method or which subtests have been used
	-	It is not clearly mentioned in this study whether the full FCE-method or which subtests have been used
2.	Objective	
	+	The objective of the study is clearly mentioned
	-	The objective of the study is not clearly mentioned
3.	Population	
	N number of subjects, G gender, A age, H health status, W work status	
	+	The 5 items N, G, A, H and W appear in the article
	+/-	3 - 4 of the 5 items appear in the article
	-	1 - 2 of the 5 items appear in the article
4.	Procedure	
	→ Intrarater Reliability	
	+	Time interval (days) between test-retest ranges from 7 to 14
	±	Time interval (days) between test-retest ranges from 3 to 6 and 15 to 21
	-	Time interval (days) between test-retest is less than 3 or more than 21
	→ Interrater Reliability	
	+	Number of raters used is more than 2
	±	Number of raters used is 2 within more than 10 measurements
	-	Number of raters used is 2 within 10 measurements or less
	→ Validity	
	+	The study design is clearly described and appears properly defined to the type of validity that it meant to be measured
	±	The study design satisfies only one of the conditions described above
	-	The study design is not clearly described and does not appear properly defined to the type of validity that it meant to be measured
5.	Statistics	
	+	The statistics used are clearly described and appear properly defined to achieve the objective of the study
	±	The study design satisfies only one of the conditions described above
	-	The statistics used are not clearly described and do not appear properly defined to achieve the objective of the study

Reliability and validity

An assessment is considered reliable when the measurements are consistent, free from significant error and repeatable over time, over the date of administration and across evaluators [26,27]. Different types of reliability are known as intrarater reliability, test–retest reliability, interrater reliability or internal consistency [22]. In this study, we looked for: (1) intrarater reliability, the consistency of measures or scores from one testing occasion to another, assuming that the characteristic being measured does not change over time, and (2) interrater reliability, the consistency of measures or score made by raters, testers or examiners on the same phenomenon [22]. As the accuracy of FCE tests is dependent on the skill of the rater, we made no distinction between intrarater reliability and test–retest reliability [28].

Table 3: The levels of reliability and validity

Level of reliability: intrarater reliability, interrater reliability and internal consistency [20,22,24]

- → Pearson Product Moment Coefficient r, Spearman Correlation Coefficient p, Somer Correlation Coefficient d*
  - high: r / p / d > 0.80
  - moderate: 0.50 ≤ r / p / d ≤ 0.80
  - low: r / p / d< 0.50
- → Intra-class Correlation Coefficient ICC
  - high: ICC > 0.90
  - moderate: 0.75 ≤ ICC ≤ 0.90
  - low: ICC < 0.75
- → Kappa value k
  - high: k > 0.60
  - moderate: 0.41 ≤ k ≤ 0.60
  - low: k ≤ 0.40
- → Cronbach's Alpha α
  - high: α > 0.80
  - moderate: 0.71 ≤ α ≤ 0.80
  - low: α ≤ 0.70
- → Percentage of agreement %
  - high: % > 0.90 and the raters can choose between more than two score levels
  - moderate: % > 0.90 and the raters can choose between two score levels
  - low: The raters can choose only between two score levels

Level of validity [20,23]

- → Face / Content validity
  - high: The test measures what it is intended to measure and all relevant components are included
  - moderate: The test measures what it is intended to measure but not all relevant components are included
  - low: The test does not measure what it is intended to measure
- → Criterion-related validity: concurrent and predictive validity
  - high: Substantial similarity between the test and the criterion measure (percentage agreement ≥ 90%, k > 0.60, r / d > 0.75)*
  - moderate: Some similarity between the test and the criterion measure (percentage agreement ≥ 70%, k ≥ 0.40, r / d ≥ 0.50)*
  - low: Little or no similarity between the test and the criterion measure (percentage agreement < 70%, k < 0.40, r / d < 0.50)*
- → Construct validity: convergent and divergent validity
  - high: Good ability to differentiate between groups or interventions, or good convergence / divergence between similar tests (r ≥ 0.60)
  - moderate: Moderate ability to differentiate between groups or interventions, or moderate convergence / divergence between similar tests (r ≥ 0.30)
  - low: Poor ability to differentiate between groups or interventions, or low convergence / divergence between similar tests (r < 0.30)

* Somer Correlation Coefficient (d) was ranged by the authors as the Pearson Product Moment Coefficient (r) and Spearman Correlation Coefficient (p)

Validity refers to the accuracy of the evaluation: an assessment is considered valid if it measures what it intends to measure and if it meets certain criterion [17,23,26,29] . In this study, we looked for: (1) face validity, the degree that a test appears to measure what it attends to measure and it is considered a plausible method to do so, (2) content validity, the degree that test items seem to be related to the construct which the test is intended to measure, (3) criterion-related validity (concurrent and predictive validity), the degree that a test is well correlated with another valued measure that has already been established being valid, and (4) construct validity (convergent and discriminant/divergent validity), the degree that a test is well correlated with a hypothetical construct or theoretical expectation [23].

To evaluate the reliability and validity levels given in each study, we defined, as for the methodological quality appraisal, a scale based on several studies (Table 3) [20,22-24] . These reliability and validity levels are expressed through different statistics as correlation coefficients (Pearson correlation coefficient, r, Spearman correlation coefficient p, Somer correlation coefficient d, Intraclass correlation coefficient, ICC, kappa value, k, Cronbach's alpha, α, percentage of agreement, %. Following our scale, we can then evaluate, for both reliability and validity, whether the FCE method used in a study has a good, moderate or poor level of reliability and/or validity.

3.3 Results

Literature search

A total of 146 potentially relevant citations were retrieved from our literature search of the five databases. Between them, 69 duplicates were identified, thus 77 references remained. The application of the inclusion criteria on their titles and abstracts (step 1) for eligibility eliminated 47 articles: one study was not written in English, French or Dutch (2%), 45 studies did not use one of the four FCEs (96%) and one study did not provide information on the reliability or validity of these FCEs (2%).

Of the remaining 30 abstracts, we read the full text and applied the inclusion criteria (step 2). Ten articles were excluded: one was not written in English, French or Dutch (10%), five did not use one of the four FCEs (50%) and four did not provide information on the reliability or validity of these FCEs (40%).

Twenty articles remained after applying the inclusion criteria on full text: 14 original papers [30-43] , and six reviews [17,29,44-47]. No article was found from the search in the database Picarta for Dutch literature. From the bibliography screening of the reviews and original papers, no more relevant articles were identified or included after applying the inclusion criteria on the full text. Therefore, 14 original articles were included in this study. Agreement between the two reviewers on the inclusion of articles was excellent (100%).

Table 4: The results of the methodological quality appraisal and the overall methodological quality

Authors	FCE method	Objective	Population	Procedure	Statistics	Methodological Quality
Brouwer S et al. (31)	+	+	+	+	+	High
Dusik LA et al. (32)	+	+	+/-	+/-	+	Moderate
Gross DP and Battié MC (33)	+	+	+	+/-	+	High
Gross DP and Battié MC (34)	+	+	+	+	+/-	High
Gross DP and Battié MC (35)	+	+	+	+	+	High
IJmker et al. (36)	+	+	+	+	+/-	High
Isernhagen SJ et al. (37)	+	+	+/-	+	+	High
Matheson LN et al. (38)	+	+	+/-	+	+	High
Reneman MF et al. (40)	+	+	+/-	+/-	+/-	Moderate
Reneman MF et al. (41)	+	+	+	-	+	Moderate
Reneman MF et al. (42)	+	+	+	+	+/-	High
Rustenburg G et al. (43)	+	+	+	+/-	+/-	Moderate

Methodological quality appraisal

During the methodological quality appraisal, two of the 14 papers were excluded. Boadella et al. [30] did not examine the intra- or interrater reliability but the reliability of the EWS in terms of learning, intensity and time of day effects. Furthermore, the study of Reneman et al. [39] on the ecological validity of the IWS was excluded because it did not discuss face, content, criterion-related or construct validity.

Therefore, the methodological quality appraisal was applied to 12 original studies. The level of agreement between reviewers in assessing the quality appraisal was excellent (100%). Table 4 provides an overview of each feature's scores of these articles. Based on the results of the methodological quality appraisal, eight articles were ranked as high [31,33-38,42], and four as moderate [32,40,41,43].

Moderate methodological quality: Four studies were evaluated as moderate concerning their methodological quality (Table 4). Two of them did not completely define the study population [32,40]. For all of them, we did not find that high quality procedures were used to achieve their objectives: three were scored as moderate [32,40,43] and one as low [41]. Concerning the concurrent validity of the EWS, the FCE outcomes were compared with the ones of other assessments but no information was provided on the reliability and validity levels of theses assessments [32]. Concerning the concurrent validity of the EWS and EK, the time interval between assessments on both FCEs was considered too long [43]. Concerning the intrarater reliability studies of the IWS, the time interval between test and retest was too short or too long [40,41].

High methodological quality: Eight studies were evaluated as high concerning their methodology quality: three studies on the intrarater and / or interrater reliability of the IWS [31,33,37], one on the concurrent validity of the IWS and EK [36] and four on the predictive and concurrent validity of the IWS [34,35,38,42].

Included studies

Tables 5 and 6 show the characteristics of all 12 included articles identified after our systematic literature search. Table 5 describes the studies on reliability and Table 6 displays those on validity.

Table 5: An overview of the included studies on the reliability of the four FCE methods

FCE method (Subtests)	Objective: Type(s) of reliability	Population (N number of subjects / G gender / A age H health status / W work status)	Procedure	Outcomes	Authors / Year of publication
Isernhagen WS 28 tests	Intrarater reliability (test-retest)	N: 30 subjects G: 24 males / 6 females A: 40 years H: chronic low back pain W: 15 out of work / 15 working	Time interval: 2 weeks	.75 ≤ ICC ≤ .87	Brouwer S et al. (31) 2003
Isernhagen WS Floor to waist lift Waist to overhead lift Horizontal lift Front carry Right/Left side carry	(1) Interrater reliability (2) Test-Retest reliability	N: 28 subjects G: 71% male / 29% female A: 41 years H: low back pain W: not working	(1) 3 raters used (2) Time interval: 2 to 4 treatment days	(1) All ICC ≥ .95 (2) All ICC ≥ .78	Gross DP and Battié MC (33) 2001
Isernhagen WS Floor to waist lift Horizontal carry Waist to crown lift	Interrater reliability	N: 3 subjects G: 3 males A: ? H: disabled for lifting W: working conditioning program	12 raters used 8 physical therapists 3 occupational therapists 1 non-clinical healthcare professional	(1) Judging lifting as light, moderate or heavy k = .68 (2) Judging lifting as light or heavy k = .81	Isernhagen SJ et al. (37) 1999
Isernhagen WS Lifting low / high Short carry Long carry two hands Long carry right hand Long carry left hand	(1) Interrater reliability (2) Intrarater reliability	N: 4 subjects G: 2 males / 2 females A: 20-30 years H: healthy W: ?	(1) 5 raters used: 3 physical therapists 2 occupational therapists (2) Time interval: 1 week to 2 months	(1) Session 1: %agreement ≥ 93% Session 2: %agreement ≥ 87% (2) % agreement ≥ .93	Reneman MF et al. (40) 2002
Isernhagen WS (1) Lifting low (2) Lifting overhead (3) Short carry	Test-Retest reliability	N: 50 subjects G: 39 males / 11 females A: 38.8 years H: chronic Low Back Pain W: 19 not working	Time interval: 1 day	(1) ICC = .87 (2) ICC = .87 (3) ICC = .77	Reneman et al. (41) 2002

ICC, Intra-class Correlation Coefficient; k, Kappa value; %, percentage of agreement

Table 6: An overview of the included studies on the validity of the four FCE methods

FCE method (Subtests)	Objective: type(s) of validity	Population (N number of subjects / G gender / A age H health status / W work status)	Procedure	Outcomes	Authors / Year of publication
Ergos WS Strength Climb/balance, Body dexterity, Reach, Talking/Hearing/Seeing	Concurrent validity	N: 70 subjects G: 70 males A: 45.1 years H: lower and upper extremities disability W: ?	(1) Ergos vs RTPE (2) Ergos vs SHOP (3) Ergos vs Valpar	(1) k = .629 for overall $.45 \leq r \leq .87$ for strength variables (2) k = .407 (3) $k \leq .45$	Dusik LA et al. (32) 1993
Isernhagen WS 3 lifting tests 3 carrying tests	Construct validity	N: 321 subjects G: 72% male/ 28% female A: 42 years H: low back injuries W: not working	Cross sectional study comparison between: (1) IWS assessments and PDI (2) IWS assessments and Pain VAS	(1) r = -.51 (2) r = -.45	Gross DP and Battié MC (34) 2003
Isernhagen WS Lifting, carrying, pushing, pulling…	Predictive validity (safely return to work)	N: 226 subjects G: 71% male/ 29% female A: 41 years H: low back injuries W: 69% of subjects working	Retrospective cohort study: ability of IWS to predict recovery	No association between IWS and recovery	Gross DP and Battié MC (35) 2004
Isernhagen WS Waist-to-overhead lift WOL Ergo-Kit Upper lifting strength ULS Upper lifting endurance ULE	Concurrent validity	N: 71 subjects G: 35 males / 36 females A: 23 years H: healthy W: students	Subsequently assessments of WOL, ULS and ULE	r = .72	IJmker et al. (36) 2003
Isernhagen WS 3 Lifting capacity tests 2 Grip force tests	Predictive validity (return to work)	N: 650 subjects (G1: 349 / G2: 301) G: G1:59.3% male/ G2:61.2% male A: G1: 40.1 years / G2: 43.1 years H: ? W: not working	Retrospective study: comparison between FCE performances of group G1 'return to work' and group G2 'not return to work'	ANOVA: differences between both groups significant at $P < 0.05$ for return to work	Matheson LN et al. (38) 2002
Isernhagen WS 14 Aactivities performed	Concurrent validity	N: 64 subjects G: 54 males / 10 females A: 38.0 years H: chronic low back pain W: 95% of subjects working	(1) IWS vs RMDQ (2) IWS vs OBPDS (3) IWS vs QBPDS	(1) p = -.17 & -.20 / d = .03 (2) $-.08 \leq d \leq .23$ (3) $-.52 \leq p \leq -.27$ $-.15 \leq d \leq .05$	Reneman MF et al. (42) 2002
Ergos WS 4 static and 6 dynamic lifting tests Ergo-Kit 4 lifting tests	Concurrent validity	N: 25 subjects G: 25 males A: 34.8 years H: healthy W: fire fighters	Time interval of 7 days between assessments on EWS and EK (order FCE counter balanced)	$.49 \leq p \leq .66$	Rustenburg et al. (43) 2004

RTPE, Rehabilitation Therapy Physical Evaluation; PDI, Pain Disability Index; VAS, Visual Analogue Scale; RMDQ, Rolland Morris Disability questionnaire; OBPDS, Oswestry Back Pain Disability Scale; QBPDS, Quebec Back Pain Disability Scale; vs, versus; k, Kappa value; r, Pearson Correlation Coefficient; p, Spearman's Rank Correlation; d, Somer's coefficient

Blankenship System:

No study was found on the reliability and validity of the Blankenship System.

Ergos Work Simulator (EWS)

The systematic literature search did not retrieve any study on the reliability of the EWS. Two studies were found on the validity of the EW [32,43]. Dusik et al. [32] examined the concurrent validity between the EWS and three other functional capacity assessments: the rehabilitation therapy physical evaluation (RTPE), the SHOP tasks and the VALPAR work sample tests. They used 70 male subjects to compare the different strength variable scores
obtained with all four assessments. The degree of concurrent validity was given by a kappa coefficient. The authors found that the EWS correlated well with the RTPE ($k=0.63$) but poorly with the SHOP and VALPAR ($k<0.45$). According to our scale (Table 4), the level of concurrent validity of the EWS is high with the RTPE and moderate with the SHOP and VALPAR. Rustenburg et al.[43] examined the concurrent validity of the EWS and the EK. Twenty-five fire fighters were assessed on the EWS and EK during lifting tests and the correlations between the two FCEs, expressed as a Spearman's Rank Correlation, varied between 0.49 and 0.66. Therefore, the concurrent validity is rated as low to moderate between the EWS and EK.

Ergo-Kit:

No study was found on the reliability of the Ergo-Kit. Two studies were found on the concurrent validity of the EK: one study on the concurrent validity of the EK and the EWS (see EWS) [43] and one on the concurrent validity of the EK and the IWS [36]. In this study, IJmker et al. [36] used 71 healthy subjects to compare the results of lifting tests of the IWS and EK. The degree of concurrent validity was expressed using a Pearson product–moment correlation and rated as moderate according to our quality appraisal scale ($r = 0.72$).

Isernhagen Work System (IWS):

The systematic literature search retrieved ten articles involving the IWS: five examined its reliability and five its validity. In these five reliability studies, four outcomes concerning the intrarater (test–retest) reliability were presented [31,33,40,41], and three outcomes about the interrater reliability [33,37,40].

Four studies evaluated the intrarater reliability (test–retest) of the IWS. Brouwer et al.[31] used 30 patients with chronic low back pain to determine the intrarater (test–retest) reliability of the whole IWS protocol (28 tests). The intrarater (test–retest) reliability was quantified with an intraclass correlation coefficient that was rated as moderate ($0.75 \leq ICC \leq 0.87$). Gross and Battié [33] used six different subtests of the IWS to determine the intrarater reliability for 28 subjects with low back pain. The intrarater reliability level was rated as moderate (all $ICC \geq 0.78$). Reneman et al. [40,41] also determined the intrarater reliability of carrying and lifting tests in healthy (n = 4) and disabled (n = 50) subjects and expressed the level of reliability with a percentage of agreement [39] and an intraclass correlation coefficient [40] that were, respectively, rated as high (% more than 93% for healthy subjects) and moderate (ICC ranged from 0.77 to 0.87 for disabled subjects) according to our scale (Table 4). To evaluate intrarater reliability, it is important to choose an optimal time interval between test and retest. This last one must not be too short, to avoid fatigue, memory or learning effects, and not too long, to avoid genuine changes in performance [26,48]. In any event, examining critically the time interval used between test and retest in three of these four studies, it should be concluded that no study used a proper and optimal procedure to evaluate the intrarater reliability. Thus, no definitive conclusion on the level of intrarater reliability of the IWS could draw from these studies.

Three studies evaluated the interrater reliability of the IWS. Gross and Battié [33] used six different subtests of the IWS to determine the interrater reliability for 28 subjects with low back pain. The interrater reliability was quantified with an intraclass correlation coefficient, which is widely recognized as the best measure of interrater reliability [28,49,50], and was rated, according to our scale, as high (all $ICC \geq 0.95$). This result is in line with the findings reported by Isernhagen et al. [37]. They used three male disabled subjects and 12 experts to measure the interrater reliability of three tests of the IWS. The degree of interrater reliability was expressed with a Kappa coefficient and was also rated as high (k = 0.81). Reneman et al. [40] also determined the interrater reliability of carrying and lifting tests in healthy subjects (n = 4). They expressed the interrater reliability with a percentage of agreement between raters that was rated as high according to our scale, showing that five raters can reliably determine the effort level during carrying and lifting tests of the IWS.

The systematic literature search retrieved five studies on the validity of the IWS. In these five validity studies, one outcome concerns the construct validity [34], two the predictive validity [35,38] and two the concurrent validity [36,42].

IJmker et al. [36] studied the concurrent validity of the IWS and the EK and the results are reported beforehand (see EK). Reneman et al. [42] examined the concurrent validity between the IWS and three self-report disability questionnaires (RMDQ, OBPDS and QBPDS). They used 64 subjects with chronic low back pain to compare the outcomes of these four assessments. The degree of concurrent validity was given by different correlation coefficients (Spearman and Somer) that were rated as low according to our scale. Gross and Battié [35] examined the predictive validity of the IWS for safe return to work using 226 patients with low back complaints. With a retrospective cohort study, the authors concluded that the predictive validity of the IWS for safe return to work was not supported. Matheson et al. [38] determined the predictive validity for return to work of five tests (three lifting capacity tests and two grip force tests) for 650 subjects with functional limitations. Using a retrospective design, they compared the test performances on the IWS between people who did return to work and those who did not. For each test, the group that returned to work (n = 349) performed better on the test than those who did not return to work (n = 301). The authors reported that the lifting and grip tests could predict return to work ($P < 0.05$). However, this study does not mention any information on the sensitivity and specificity of the measures used to predict return to work. Gross and Battié [34] used 321 patients with low back complaints to evaluate the construct validity of the IWS and both the Pain Disability Index (PDI) and a pain visual analogue scale (VAS). The correlations of the IWS and the PDI ($r = 0.51$) and the VAS ($r = 0.45$) were rated as low to moderate, showing that the IWS is poorly related to these pain rating scales.

3.4 Discussion

In the present systematic literature search, we tried to identify the available evidence in the literature on the reliability and validity of four FCEs: BS, EWS, EK and IWS. To retrieve relevant literature, we used different electronic databases (CINAHL, Medline, Embase, OSH-ROM and Picarta) and combined synonyms of functional capacity evaluation with synonyms of reliability and validity. After the search in the electronic databases and the application of the inclusion criteria, 14 original articles were included. From these studies, one study was excluded as it did not evaluate one of reliability types we were looking for, and one examining the ecological validity of the IWS was also excluded as this form of validity appears not clearly defined. Then, we finally included 12 original articles: one concerning the validity of the EWS, one concerning the concurrent validity of the EWS with the EK, one concerning the concurrent validity of the EK with the IWS, five concerning the reliability of

the IWS and four concerning its validity. No study concerning the reliability and validity of the BS, EWS and EK was retrieved from the literature.

While a systematic search of the literature was performed, there may be a few potential limitations of our review concerning the included articles. Even if we tried to identify all relevant articles, there can be potential relevant articles that were omitted as other articles may have used other keywords than the ones we defined and used in our literature search. Other articles may also be written in languages other than English, Dutch or French. However, considering the large definition of the keywords and databases, we are in the opinion that the most relevant articles on the reliability or validity of these FCEs should have been identified and selected from our systematic literature search or from the bibliography screening of the reviews or original papers. Our systematic literature search allows us to conclude that studies on the reliability and validity of the BS, EWS and EK are lacking. Concerning the IWS, several authors studied its intrarater and interrater reliability, and its construct, concurrent and predictive validity. The interrater reliability and the predictive validity of the IWS have been evaluated as moderate to good, while the procedures of the intrarater reliability studies were not considered rigorous enough to draw any conclusion. The construct and concurrent validity of the IWS were not demonstrated.

For any kind of test or measurement, scientific acceptance should be achieved: reliability and validity should be demonstrated. Overall, five issues must be addressed in the selection and use of any functional test: safety, reliability, validity, practicality and utility [51]. This hierarchy requires that each of the factors must be addressed so that the factors which are presented earlier are maintained: demonstration of acceptable reliability is a precursor for demonstrating an instrument's validity [28,48]. If an FCE measurement is not reliable, tests results are not consistent and it would be thus impossible to demonstrate its validity [17]. Therefore, any study concerning the validity of one of the four FCEs should refer to or mention its reliability. Dusik et al. [32], IJmker et al. [36] and Rustenburg et al. [43] examined the concurrent validity of the EWS and the EK without referring to any reliability study: no level of reliability of the EWS and EK could be found. Regarding the studies on the validity of the IWS [34,35,38,42], all authors did mention its level of reliability and refer to the studies in their bibliography.

''Concurrent validity'' is defined as the correlation of a (new) instrument with a criterion called 'gold standard', that is already established and assumed reliable and valid [27,28]. In the

studies of Dusik et al. [32], IJmker et al. [36] and Rustenburg et al. [43], the use of the term concurrent validity appears inappropriate, as no gold standard is available. Therefore, it would have been more suitable and pertinent to talk about a comparison or correlation study instead of a concurrent validity study. Furthermore, in a concurrent validity study, both measures (instrument and gold standard) should be performed at the same point of time, thus concurrently, so to reflect the same behaviour [26-28]. In their studies, Dusik et al. [32] and Rustenburg et al. [43] did not assess the different assessment methods at the same point of time (concurrently), making their reference as concurrent validity studies even less suitable.

Functional capacity evaluations are principally used in rehabilitation and work disability. In a rehabilitation context, physical therapists try to improve the physical abilities of patients who suffer from musculoskeletal injuries and disease. They generally use an FCE as an instrument to evaluate a rehabilitation program or a treatment by measuring the physical abilities of patients before and after this rehabilitation program. They use FCE as a periodic examination to modify the treatment if necessary and to develop a (new) rehabilitation strategy adapted to the current physical abilities of the patient. From the FCE test results and their personal judgment and diagnosis, physical therapists will decide whether a patient could reintegrate into the community or workplace after injury or illness. In work disability, FCEs are used by occupational therapists, insurance companies or rehabilitation counselors to help people suffering from injuries or disease and to improve their ability to perform tasks in their working environment. FCE test results are used to evaluate whether an injured worker can work and when he can return to work. Furthermore, during a work disability claim, insurance entities use FCEs to evaluate the percentage of work loss of an injured worker to determine his work disability compensation. Thus, FCE test results can have large financial consequences not only for the worker and his family, but also for governments and insurance entities. As our systematic literature review showed, reliability and validity of the BS, EWS and EK have not been demonstrated yet. For the IWS, reliability is good. Therefore, we should be prudent with the use of one of these FCE test results in rehabilitation and work disability, especially in claim procedures.

Although FCE methods such as the IWS look promising, and knowing that FCEs are used mainly in rehabilitation and work disability to evaluate the physical abilities of disabled people, more studies are needed to demonstrate the reliability and the validity of these FCEs, using especially disabled subjects. These studies should also concentrate on the definition and

selection of appropriate procedure in order to increase their methodological quality, allowing then to conclude objectively on the reliability and validity of the BS, EWS, EK and IWS.

Reference List

1. Ferrari R, Russel AS (2003) Regional musculoskeletal conditions. Best Pract Res ClinRheumatol 17: 57-70
2. Marras WS (2000) Occupational low back disorder causation and control. Ergonomics 43: 880-902
3. Palmer KT (2003) Pain in forearm, wrist and hand. Best Pract Res Clin Rheumatol 17: 113-135
4. Reginster JY (2002) The prevalence and burden of arthritis. Rheumatology 41(Suppl 1): 3-6
5. IEBS Institute for Employee Benefit Schemes (UWV) (2001). Work disability development: annual survey. [Uitvoering Werknemers Verzekeringen. Ontwikkeling arbeidsongeschiktheid: Jaaroverzicht WAO, WAZ en Wajong: in Dutch]
6. Statistic Central Desk. http://www.cbs.nl [Centraal Bureau voor Statistiek: in Dutch]
7. Strong S (2002) Functional capacity evaluation: the good, the bad and the ugly. Occup Ther Now 5-9
8. Tuckwell NL, Straker L, Barrett TE (2002) Test–retest reliability on nine tasks of the physical work performance evaluation. Work 19: 243-253
9. Vasudevan SV (1996) Role of functional capacity assessment in disability evaluation. J Back Musculoskelet 6: 237-248
10. Lechner DE (1998) Functional capacity evaluation. In: King PM (ed) Sourcebook of occupational rehabilitation. Plenum, New York pp 209-227
11. Harten JA (1998) Functional capacity evaluation. Occup Med State Art Rev 13: 209-212
12. Innes E, Straker L (2002) Workplace assessments and functional capacity evaluations: current practices of therapists in Australia. Work 18: 51-66
13. Blankenship KL (1994) The Blankenship system functional capacity evaluation: the procedure manual. The Blankenship Corporation, Macon
14. EWS FCE: Ergos Work Simulator. Users Guide. Work Recovery System Inc., Tucson, Arizona
15. EKFCE Ergo-Kit functional capacity evaluation (2002) User manual. Enschede, The Netherlands: Ergo Control. [Ergo-Kit Functionele Capaciteit Evaluatie. Handleiding: in Dutch]

16. IWSFCE: Isernhagen Work System Functional Capacity Evaluation. Manual. Duluth, Minnesota, USA
17. King PM, Tuckwell N, Barrett TE (1998) A critical review of functional capacity evaluations. Phys Ther 78: 852-866
18. Lindeboom D, Bachoe S, Karsemeijer E, Faber L (2003) The place of FCE in the area of investigation of work-related problems, rehabilitation and integration. [De plaats van FCE op gebied van onderzoek naar arbeidsgebonden problematiek, revalidatie en integratie. Rapport Arbeidsreïntegratie, Hulpmiddelen en Ergonomie: in Dutch]
19. Gardener L, McKenna K (1999) Reliability of occupational therapists in determining safe, maximal lifting capacity. Aust Occup Ther J 46: 110-119
20. Altman DG (1991) Practical statistics for medical research. Chapman and Hall, London
21. Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures. Control Clin Trials 12: 142S-158S
22. Innes E, Straker L(1999) Reliability of work-related assessments. Work 13: 107-124
23. Innes E, Straker L (1999) Validity of work-related assessments. Work 13: 125-152
24. Nunnally JC (1978) Psychometric theory, 2nd edn. McGraw-Hill, New York
25. Weiner EA, Stewart BJ (1984) Assessing individuals. Little Brown, Boston
26. Carmines EG, Zeller A (1979) Reliability and validity assessment. Sage Publications, Iowa
27. Streiner DL, Norman GR (2003) Health measurement scales. Oxford University Press, New York
28. Portney LG, Watkins MP (2000) Foundations of clinical research: applications to practice. Appleton and Lange, Norwalk
29. Schultz-Jonhson K (2002) Functional capacity evaluation following flexor tendon injury. Hand Surg 7: 109-137
30. Boadella JM, Sluiter JK, Frings-Dresen MHW (2003) Reliability of upper extremity tests measured by the ErgosTM work simulator: a pilot study. J Occup Rehabil 13: 219-232
31. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JMH, Goëken LNH (2003) Test–retest reliability of the Isernhagen Work Systems functional capacity evaluation in patients with chronic low back pain. J Occup Rehabil 13: 207-218
32. Dusik LA, Menard MR, Cooke C, Fairburn SM, Beach GN (1993) Concurrent validity of the ERGOS work simulator versus conventional functional capacity evaluation techniques in a workers' compensation population. J Occup Med 35: 759-767

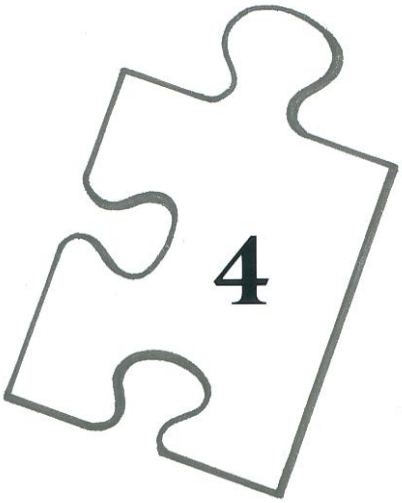
33. Gross DP, Battié MC (2001) Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther 82: 364-371
34. Gross DP, Battié MC (2003) Construct validity of kinesiophysical functional capacity evaluation administrated within a worker's compensation environment. J Occup Rehabil 13: 287-295
35. Gross DP, Battié MC (2004) The prognostic value of functional capacity evaluation in patients with chronic low back pain: Part 2. Spine 29: 920-924
36. IJmker S, Gerrits EHJ, Reneman MF (2003) Upper lifting performance of healthy young adults in functional capacity evaluations: a comparison of two protocols. J Occup Rehabil 13: 297-305
37. Isernhagen SJ, Hart DL, Matheson LM (1999) Reliability of independent observer judgements of level of lift effort in kinesiophysical functional capacity evaluation. Work 12: 145-150
38. Matheson LN, Isernhagen SJ, Hart DL (2002) Relationships among lifting ability, grip force, and return to work. Phys Ther 82: 249-256
39. Reneman MF, Joling CI, Soer EL, Goëken LNH (2001) Functional capacity evaluation: ecological validity of three static endurance tests. Work 16: 227-234
40. Reneman MF, Jaegers SMHJ, Westmaas M, Goëken LNH (2002) The reliability of determining effort level of lifting and carrying in a functional capacity evaluation. Work 18: 23-27
41. Reneman MF, Dijkstra PU, Westmaas M, Goëken LNH (2002) Test–Retest reliability of lifting and carrying in a 2-day functional capacity evaluation. J Occup Rehabil 12: 269-275
42. Reneman MF, Jorritsma W, Schellekens JMH, Goëken LNH (2002) Concurrent validity of questionnaire and performancebased disability measurements in patients with chronic nonspecific low back pain. J Occup Rehabil 12: 119-129
43. Rustenburg G, Kuijer PPFM, Frings-Dresen MHW (2004) The concurrent validity of the ERGOS work simulator and the Ergo-Kit with respect to maximum lifting capacity. J Occup Rehabil 14: 107-118
44. Gibson L, Strong J (1997) A review of functional capacity evaluation practice. Work 9: 3-11
45. Jones T, Kumar S (2003) Functional capacity evaluation of manual materials handlers: a review. Disabil Rehabil 25: 179-191

46. Lechner DE (2002) The role of functional capacity evaluation in management of foot and ankle dysfunction. Foot Ankle Clin N Am 7: 449-476
47. Tramposh AK (1992) The functional capacity evaluation: measuring maximal work abilities. Occup Med State Art 7: 113-124
48. Matheson LN, Mooney , Grant JE, Legget S, Kenny K (1996) Standarized evaluation of work capacity. J Back Musculoskelet 6: 249-264
49. Fleiss JL (1986) The design and analysis of clinical experiments. Wiley, New York
50. Tinsley HEA, Weiss DJ (1975) Interrater reliability and agreement of subjective judgements. J Couns Psychol 22: 358-376
51. Hart DL, Isernhagen SJ, Matheson LN (1993) Guidelines for functional capacity evaluation of people with medical conditions. J Orthop Sport Phys 18: 682-686

Chapter 4

Reliability and agreement of five Ergo-Kit FCE lifting tests in subjects with low back pain

Vincent Gouttebarge, Haije Wind, P. Paul F. M. Kuijer, Judith K. Sluiter, Monique H.W. Frings-Dresen. Arch Phys Med Rehabil 2006;87:1365-1370.



Abstract

Objectives

To assess the interrater reliability and agreement of five Ergo-Kit (EK) FCE lifting tests in subjects with low back pain.

Methods

A within-subjects design was used to assess on two occasions, t1 and t2, by two different raters, R1 and R2, twenty-four subjects with low back pain (10 males and 14 females) on five EK lifting tests (two isometric and three dynamic). The time interval between both test sessions was three days. Interrater reliability level was expressed with Intra-Class Correlation Coefficient (ICC) and level of agreement between raters with Standard Error of Measurement (SEM).

Results

ICCs means (reliability) of isometric and dynamic EK lifting tests ranged from 0.94 to 0.97, and SEMs values (agreement) ranged from 1.9 to 8.6 kg.

Conclusions

There was good reliability and agreement between raters of the isometric and dynamic EK lifting tests in subjects with low back pain, which support the use of these tests in practice to assess functional lifting capacity.

4.1 Introduction

Low back pain (LBP) is recognized as a major public health problem throughout the world. In fact, it is the most common and most costly musculoskeletal disorder in all industrialized countries [1-4]. The sickness-related absences and work disability claims resulting from LBP place a tremendous financial strain on patients and their communities [5-10]. Given this condition's social impact and its financial ramifications for society, professionals in work disability and rehabilitation care need clinical instruments to accurately assess the functional capacity of LBP patients and thus enhance the effectiveness of the return to work process.

Clinical instruments are principally used to measure relevant changes in people over time [11]. The purpose of functional capacity evaluation (FCE) methods is to provide comprehensive, performance-based assessments that measure the current functional physical abilities of people with musculoskeletal complaints [12-16]. In the Netherlands, the Ergo-Kit is an FCE method devised to report the functional physical capacity of workers. The Ergo-Kit relies on a battery of standardized tests that reflect work-related activities such as standing, walking, lifting, carrying, and reaching [17].

As with any clinical test or instrument, the clinimetric properties of the Ergo-Kit must be defined before it can be legitimately applied for discriminative or evaluative purposes [18-20]. Clinimetric properties, also referred to as psychometric properties [21], reflect the quality of clinical measurements and are crucial in selecting and using instruments— either for clinical practice or research [22,23]. Several studies have shown that despite their use in both evaluative and discriminative settings, there is little information currently available about the clinimetric properties of FCE methods (including the Ergo-Kit), such as their reproducibility, reliability, responsiveness, and validity [16,24-29]. Reproducibility is a major quality indicator [26,30,31] and relates to 2 concepts that are not always differentiated from each other: reliability and agreement [23]. Reliability refers to the test's ability to distinguish one subject from another despite any measurement errors. Agreement, on the other hand, concerns the absolute measurement error, as it evaluates how close the scores are in repeated measurements [23,32].

Reproducibility studies should address populations that are relevant to the implementation of tests or instruments in the field [33]. The reliability of the Ergo-Kit isometric and dynamic lifting tests has been assessed in adults with no musculoskeletal complaints [34], but they should also be evaluated in subjects who do report these complaints. It is also important to establish

interrater reliability and agreement to ensure adequate and meaningful interpretation of variations in the test measurements of different raters [35]. In this study, we evaluated the reproducibility (i.e., reliability and agreement between raters) of the Ergo-Kit isometric and dynamic lifting tests in subjects with LBP.

4.2 Methods

Participants

Fourteen physiotherapy (PT) centers in the southern section of Amsterdam were contacted for permission to recruit patients from their practices. All patients were initially contacted by their physiotherapists, who briefly explained the experimental procedures. The patients interested in participating received a folder containing detailed information on the study protocol and were asked to contact the first author (VG). Participant eligibility was determined through telephone interviews, during which potential subjects were asked several questions that were intended to determine whether the subjects met the 3 inclusion criteria: (1) age between 18 and 65 years, (2) had LBP in the last 3 months, and (3) because of LBP, had limited physical capacity in daily activities at home and at work. We defined LBP as 1 or more episodes of pain or stiffness in the low back area within the past 3 months that lasted for a minimum of 7 consecutive days. A power analysis (confidence interval [CI] method with confidence level of .95, correlation coefficient sets at .90 and limit at .80) indicated that 23 subjects were required for the study. Prior to enrollment, subjects received verbal and written information on the study procedures and signed statements of informed consent. In addition, the subjects were free to withdraw from the study at any time. The study was performed in accordance with the Declaration of Helsinki and was approved by the Medical Ethics Committee of the Academic Medical Centre in Amsterdam.

Ergo-Kit tests: selection, description and outcomes

The standard protocol of the Ergo-Kit assesses 55 subtests, and takes approximately 3 hours to complete. Of the 7 Ergo-Kit physical agility tests concerned with manipulation, balance, strength, and endurance tests that are associated with musculoskeletal complaints, 2 have been shown to be unreliable in adults without musculoskeletal complaints [34]. Consequently, we used only 5 of the lifting tests in this study (fig 1). Two were isometric lifting tests: a back-torso lift test (BTLT) and a shoulder lift test (SLT). The other 3 were dynamic lifting tests: carrying lifting strength test (CLST), lower lifting strength test (LLST), and upper lifting strength test (ULST). Table 1 presents Ergo-Kit lifting test descriptions and outcomes.

Standardized procedures were performed as described in the Ergo-Kit handbook [17]. The Ergo-Kit protocol normally includes 2 tests on the Jamar hand dynamometer, 4 reach tests, and 5 manipulation tests between these 5 lifting tests. The testing order for the 5 lifting tests was not modified.

Table 1: EK test descriptions and outcomes [17]

EK® Test	Description	Outcome
Back-torso lift test (Btlt) Shoulder lift test (Slt)	Use of a "back and leg dynamometer" fixed on a platform, a chain and a handle. Handle is set at patella height for BTLT (fig 1A) and at elbow height for SLT (fig 1B). Maximal pulling during 4 s, 2 tries per test.	Maximal isometric lift capacity (kg)
Carrying lifting strength test. (Clst) Lower lifting strength test (Llst) Upper lifting strength test (Ulst).	Use of a stand with two vertically adjustable shelves, a box with different weights and a step (20cm). Following standardized procedure, weight is added to the box (2.5, 5, 7.5 or 10 kg), depending on the subject's coordination in the task, subject's perception of the weight of the box, and subject complaints. 4-6 carries 5 m for CLST (fig 1C), 4-6 lifts from knuckle height to step for LLST (fig 1D) and 4-6 lifts from knuckle to acromion height for ULST (fig 1E).	Maximal safe weight for lifting (kg)

Raters

A list of the 32 available raters in the Netherlands who were certified for Ergo-Kit assessment was obtained from the provider of this FCE method. All had completed the same training program, which consisted of 4 instruction days and at least 12 hours of practice. Because the test assessments were to take place in Amsterdam, selection was limited—for practical reasons— to raters who worked within a 40-km radius of the city. This left 3 raters, 2 of whom were selected at random and agreed to participate. Both raters (R1, R2) had between 4 and 5 years of experience performing the assessments. The raters received financial compensation and travel reimbursement for their participation.

Figure 1: Five Ergo Kit lifting tests: (a) Back-torso lift test (Btlt), (b) Shoulder lift test (Slt), (c) Carrying lifting strength test (Clst), (d) Lower lifting strength test (Llst) and (e) Upper lifting strength test (Ulst).

(a)

(b)

(c)

(d)

(e)

Procedure

We used a within-subjects design to assess reliability and agreement. Each subject was assessed at 2 different times (t1, t2) by the 2 different raters (R1, R2). Raters assessed all subjects independently and were blinded to the other's test results. The time interval between t1 and t2 was set at 3 days, as this was considered sufficient to prevent carry-over effects and to give subjects time to recover from the first assessment [36-38]. In addition, each subject was assessed at the same time of the day [39]. Subjects were divided into 2 groups, based on their availability, and the raters assessed both groups in counterbalanced order: 1 subject group was assessed at t1 by R1 and at t2 by R2, and 1 group at t1 by R2 and at t2 by R1. Prior to the second assessment, all subjects were asked whether they had recovered satisfactorily from the first assessment. If they had not, they were not allowed to undergo the second assessment, but were permitted to participate in 2 new test sessions at a later date.

Low Back Pain: pain intensity and disability

Before both assessments, the patients were asked to complete an existing Dutch translation of the Von Korff questionnaire [40] about their LBP and related disability. This was done to permit us to evaluate whether their health status had changed between t1 and t2. The Von Korff questionnaire has shown a moderate-to-good correlation with other self-reported disability instruments such as the Medical Outcome Study 36-Item Short-Form Health Survey and the Roland-Morris Disability Questionnaire; it has been evaluated as reliable and valid in study samples similar to the one in this study [41,42]. The Von Korff questionnaire assesses pain and disability experienced in the past 6 months; therefore in order to fit our inclusion criteria, we adjusted it to consider only a 3-month prevalence of LBP and disability. Current pain intensity was assessed with 3 questions that were scored on a scale of 0 (no pain) to 10 (the worst pain possible). Disability due to LBP was assessed with 4 questions about the number of days the subjects were disabled and their ability to perform activities and/or work (scored on a scale of 0 to 10). Two total scores were calculated: a 0 to 100 pain intensity score based on the mean of the pain intensity questions multiplied by 10, and a 0 to 100 disability score based on the mean of the disability questions multiplied by 10 [40].

Kinesiophobia and Low Back Pain

Subjects were asked to fill in the Dutch version of the Tampa Scale of Kinesiophobia (TSK) [43] to assess their fear of reinjury caused by physical movement and activity. The TSK covers 17 items, each of which is scored on a 4-point Likert scale ranging from "strongly disagree" to "strongly agree." For each subject, a total score ranging from 17 to 68 was calculated after inversion of the individual scores for items 4, 8, 12, and 16. The TSK has shown good reliability and validity in different study populations [44,45]. The TSK was completed after each test session to avoid eventual effects on subjects' performance provided by the assessment of this questionnaire.

Data analysis

Means, standard deviations (SDs), and ranges were calculated for each test for raters 1 and 2. The level of reliability was expressed with an intraclass correlation coefficient (ICC) [20,46,47] and determined with the test scores assessed by the 2 raters. We used the ICC model 2.1.A, based on a mixed 2-way analysis of variance, as defined by Shrout and Fleiss [48]. The 95% CI was calculated for each ICC mean. The ICC and 95% CI values were evaluated as follows [20,26,49,50]: "low" reliability when ICC means and/or CI lower bounds were lower than 0.50;

"moderate" reliability when ICC means and/or CI lower bounds ranged from 0.50 to 0.80; and "high" reliability when ICC means and CI lower bounds were greater than 0.80. To assess the raters' stability in repeated measurements over time and to gain an insight into the clinical relevance of the Ergo-Kit lifting tests, agreement was expressed with the standard error (SE) of measurement (SE of measurement = √ [var(raters) + var(error)] or SE of measurement = SD X √[1 – ICC]) and its 95% CI (95% CI = 1.96 X SE of measurement) [20,33]. Using a general linear model, we calculated 3 different components of variation, variance between subjects (var[subjects]), variance between raters (var[raters]), and variance due to measurement error (var[error]). To explore the stability of the patients' health status from 1 test to another, their mean scores for LBP pain intensity and related disability (disability score), and kinesiophobia were calculated from the Von Korff questionnaire and TSK at t1 and t2. For these 3 variables, statistical differences between t1 and t2 were explored with paired t tests. All analyses were performed with the statistical analysis software SPSS for Windows.

4.3 Results

Participant characteristics

Twenty-five subjects with LBP (11 men, 14 women) were recruited for this study. All subjects were working either part-time or full-time in a variety of professions. One subject was not able to perform the second test assessment because he did not recover properly from the first test. The subjects' mean age ± SD was 49 ± 8 years (range, 34–63 years), their mean height was 175 cm (range, 158–195 cm), and their mean body weight was 78 kg (range, 48–97 kg). There were few differences between t1 and t2 in terms of the subjects' LBP pain intensity (p = .003) and related disability, and their subjects' TSK mean scores (table 2). These small differences in average pain intensity, average disability, and average TSK scores, however, do not appear to be clinical relevant changes within subjects from 1 test session to another.

Table 2: Mean (SD) scores and observed differences of LBP pain intensity, disability and kinesiophobia

Items	N	t1		t2		|D|	P-Values
		Mean	SD	Mean	SD		
Pain intensity, 0-100	24	62.7	19.9	56.2	19.1	6.5	.003
Disability score, 0-100	24	46.4	29.5	41.0	26.0	5.4	.141
TSK total score, 17-68	24	39.3	6.7	40.2	6.2	0.9	.369

N, number of subjects; t1, test session 1; t2, test session 2; SD, standard deviation; |D|, absolute difference between t1 and t2.

Reliability

Table 3 presents the averages, SDs, and ranges in scores for all 5 Ergo-Kit tests for both sessions assessed by the raters, their mean ICCs, and corresponding 95% CIs. The level of interrater reliability was high for both isometric strength tests (BTLT, SLT), as their mean ICCs were 0.97 and 0.96, respectively, with CI lower bounds of 0.94 and 0.91, respectively. The mean ICCs for the 3 dynamic strength tests were 0.95 for the CLST and the ULST, and 0.94 for the LLST. The corresponding CI lower bounds (0.84, 0.89, and 0.85, respectively) are considered highly reliable.

Variation component

The variation components (between subjects, between raters, systematic error) for all 5 tests are presented in table 4. Given the ratio between all 3 variation components in all 5 tests, var[raters] is relatively small, whereas var[subjects] is relatively high.

Agreement

Table 4 presents the SE of measurement and the 95% CI for each test, which offers a clear picture of the agreement between the raters. The SE of measurements, expressed in kilograms, are small, especially given the mean values of the different tests (see table 3). For instance, the BTLT mean score from both test assessments approaches 64 kg, its SE of measurement 8.6 kg, and its CI 47 to 81 kg. This indicates that an increase or decrease of 17 kg from the observed score cannot be interpreted as a change resulting from a measurement error.

Table 3: Test scores assessed by R1 and R2 (time interval of 3 days) and interrater reliability levels

Tests (kg)	N	Rater 1			Rater 2			ICC	ICC 95% CI
		Mean	SD	Range	Mean	SD	Range		Lower - Upper
Back-torso lift test	24	65.9	38.3	20.5 – 180.5	63.3	39.5	12.5 – 177.5	.97	.94 - .99
Shoulder lift test	24	37.6	18.3	14.0 - 72.0	38.9	19.1	10.0 - 76.5	.96	.91 - .98
Carrying lifting strength test	24	24.5	9.7	10.0 - 47.5	22.1	11.2	7.5 - 47.5	.95	.84 - .98
Lower lifting strength test	24	23.8	11.1	7.5 - 47.5	21.8	10.6	7.5 - 47.5	.94	.85 - .97
Upper lifting strength test	24	17.0	6.3	7.5 - 32.5	17.1	6.6	7.5 - 30.0	.95	.89 - .98

N, number of subjects; SD, standard deviation; ICC, Intra-Class correlation coefficient; CI, confidence interval; kg, kilograms; s, seconds.

Table 4: Variation components and indicators of agreement per EK test

Tests (kg)	N =	VAR[subjects]	VAR[raters]	VAR[error]	SEM	SEM 95%CI
Back-torso lift test	24	1444.90	0.36	72.66	8.6	X ± 16.7
Shoulder lift test	24	324.65	-0.16*	25.06	5.0	X ± 9.8
Carrying lifting strength test	24	101.40	2.50	8.83	3.4	X ± 6.6
Lower lifting strength test	24	105.92	1.68	11.87	3.7	X ± 7.2
Upper lifting strength test	24	37.55	-0.16*	3.93	1.9	X ± 3.8

N, number of subjects; VAR[subjects], variance between subjects; VAR[raters], variance between raters; VAR[error], error variance; SEM, Standard error of measurement; CI, confidence interval; X, observed test score; *, treat negative variance components as though they are zero.

4.4 Discussion

Our purpose in this study was to evaluate the reliability and agreement between 2 raters of 5 Ergo-Kit lifting tests in subjects with LBP. For both of the isometric lifting tests, reliability between raters was considered high, a finding that is in line with other studies [51-54]. The 3 dynamic lifting tests were also found to be highly reliable. The ICC is an accepted measure of reliability when it comes to the discriminative capacity of a test. ICC values are sensitive to the heterogeneity of the study population: when measurement error variability is small compared to the performance variability between subjects, ICC values can be high, approaching 1, as they did in this study. The SDs in all 5 tests were high, showing significant variability in test scores between subjects.

Our findings in this study are in line with reliability studies of other types of lifting tests [14,27,52,55-57], but especially of another FCE method, the Isernhagen Work System (IWS) [29,58-62]. The methods differ in their design (material used and needed) and assessment method (step-by-step test protocol to get the end point of the lifting tests). The IWS uses a kinesiophysical approach, relying on the therapist's expertise (observations) to determine maximum lifting capacity rather than on patient reports (pain, discomfort), while the Ergo-Kit is based on both the therapist's expertise and patient reports. It is possible, however, to draw comparisons of studies of both lifting tests because dynamic lifting capacity seems to be the construct being measured in both Ergo-Kit and IWS lifting tests. Gross and Battie [58], Brouwer et al. [59], and Reneman et al. [60] performed their studies also with subjects with LBP, quantified their outcomes with an ICC as well, and found high levels of intra- and interrater reliability of IWS dynamic lifting tests (ICC range 0.75–0.98). So, because dynamic lifting tests from both Ergo-Kit and IWS FCE methods are reliable, it would be relevant to assess these lifting tests concurrently with the same subjects to obtain an appropriate insight into whether they could be used interchangeably.

This study is the first to evaluate agreement between FCE tests. For agreement, different statistics are commonly used, such as the Bland-Altman visual plotting method [63], smallest real difference [64], and SE of measurement. As variations between raters were nearly nil (see table 4), it can be concluded that the variations in test scores between both assessments were not the result of disagreement between the 2 raters, but rather to performance variations within subjects. Because the Ergo-Kit tests are used in PT settings for evaluative purposes, reproducibility and responsiveness are 2 major properties that need to be evaluated [11]. In this

study, we examined the reproducibility of the tests by calculating the SEs of measurement and CIs. Some suggestions about the responsiveness of these tests may also be made, however. For instance, the Ergo-Kit tests should be able to detect clinically relevant changes within subjects during repeated evaluations throughout a rehabilitation program. Safe amount of minimal change that has to be found to conclude that change in subjects' performance is due to a real change and not to measurement error, may be found by checking the SEs of measurement and CIs. As in other studies [65,66], SEs of measurement could be expressed as a percentage of the mean test score (i.e., at t1: BTLT, 13.1%; SLT, 13.3%; CLST, 13.9%; LLST, 15.5%; ULST, 11.2%). Because no similar data of the SE of measurement and SE of measurement percentage for FCE lifting tests have been previously reported, the set-up of an SE of measurement's cutoff value for clinical relevance cannot be retrieved from literature, and should be based on the practitioner's knowledge of the lifting tests. The SE of measurement percentage values we found in this study were lower than the ones found for isokinetic strength tests [65,66], which suggests a sufficient agreement level of the Ergo-Kit lifting tests, and thus makes their clinical use legitimate.

4.5 Conclusion

Our results suggest that the reproducibility (i.e., reliability and agreement between raters) of 5 Ergo-Kit tests in subjects with LBP was good. Criterion-related and construct validity appear to be the topics that merit the most attention in future studies.

Acknowledgements: we thank Henrica C. de Vet, PhD, for her statistical advice, and the two raters who assessed the Ergo-Kit tests for their time and effort in participating in this study.

Reference list

1. Wyatt M, Underwood MR, Scheel IB, Scheel IB, Nagel P (2004) Back pain and health policy research: the what, why, how, who, and when. Spine 29: E468-E475
2. Liddle SD, Baxter GD, Gracey JH (2004) Exercise and chronic low back pain: what works? Pain 107: 176-190
3. Kääriä S, Kaila-Kangas L, Kirjonen J, Riihimaki H, Luukkonen E, Leino-Arjas P (2005) Low back pain, work absenteeism, chronic back disorders, and clinical findings in the low back as predictors of hospitalization due to low back disorders. Spine 30: 1211-1218
4. Speed C (2005) Low back pain. BMJ 328: 1119-1121
5. Elders LAM, Burdorf A (2001) Interrelations of risk factors and low back pain in scaffolders. Occup Environ Med 58: 597-603
6. IJzelenberg W, Burdorf A (2004) Impact of musculoskeletal co-morbidity of neck and upper extremities on healthcare utilisation and sickness absence for low back pain. Occup Environ Med 61: 806-810
7. Institute for Employee Benefit Schemes UWV. (2001) Statistical information on medical classification in work disability claim 1999. [Statistische informatie over medische classificatie in WAO, WAZ en Wajong 1999: in Dutch]
8. Central Bureau of Statistics. http://www.cbs.nl. [Centraal Bureau voor Statistiek: in Dutch]
9. Picavet HSJ (2004) Multimedia campaign on low back pain prevention: potential health benefits. RIVM [Een multimedia campagne gericht op de preventie van lage rugklachten: de potentiële gezondheidswinst: in Dutch]
10. Grabois M (2005) Management of chronic low back pain. Am J Phys Med Rehabil 84: S29-S41
11. Guyatt G, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 40: 171-178
12. King PM (1998) Sourcebook of occupational rehabilitation. New York: Plenum Press
13. Strong S (2002) Functional Capacity Evaluation: the good, the bad and the ugly. Occupational Therapy Now: 5-9
14. Tuckwell NL, Straker L, Barrett TE (2002) Test-retest reliability on nine tasks of the Physical Work Performance Evaluation. Work 19: 243-253
15. Vasudevan SV (1996) Role of functional capacity assessment in disability evaluation. J Back Musculoskeletal Rehabil 6: 237-248

16. Strong S, Baptiste S, Cole D, Clarke J, Costa M, Shannon H, Reardon R, Sinclair S (2004) Functional assessment of injured workers: A profile of assessor practices. Can J Occup Ther 71: 13-23
17. Ergo-Kit for functional capacity evaluation (2002) User manual. Enschede, the Netherlands: Ergo Control. [Ergo-kit Functionele Capaciteit Evaluatie. Handleiding: in Dutch]
18. Matheson LN, Mooney V, Grant JE, Legget S, Kenny K (1996) Standardized evaluation of work capacity. J Back Musculoskeletal Rehabil 6: 249-264
19. Mooney V (2002) Functional capacity evaluation. Orthopedics 25: 1094-1099
20. Portney LG, Watkins MP (2000) Foundations of clinical research: Applications to practice. New Jersey: Prentice-Hall
21. Streiner DL (2003) Clinimetrics vs. psychometrics: an unnecessary distinction. J Clin Epidemiol 56: 1142-1145
22. Feinstein AR (1987) Clinimetrics. New Haven: Yale University Press
23. De Vet HCW, Terwee CB, Bouter LM (2003) Current challenges in clinimetrics. J Clin Epidemiol 56: 1137-1141
24. Wind H, Gouttebarge V, Kuijer PPFM, Frings-Dresen MH (2005) Assessment of functional capacity of the musculoskeletal system in the context of work, daily living, and sport: a systematic review. J Occup Rehabil 15: 253-272
25. Innes E, Straker L (1999) Validity of work-related assessments. Work 13: 125-152
26. Innes E, Straker L (1999) Reliability of work-related assessments. Work 13: 107-124
27. Gardener L, McKenna K (1999) Reliability of occupational therapists in determining safe, maximal lifting capacity. Aust Occup Ther J 46: 110-119
28. Strong S, Baptiste S, Clarke J, Cole D, Costa M (2004) Use of functional capacity evaluations in workplaces and the compensation system: A report on workers' and report users' perceptions. Work 23: 67-77
29. Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2004) 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 77: 527-537
30. Hart DL, Isernhagen SJ, Matheson LN (1993) Guidelines for functional capacity evaluation of people with medical conditions. J Orthop Sports Phys Ther 18: 682-686
31. Nunnally JC. (1994) Psychometric theory. New York: McGraw-Hill

32. De Vet HCW. (1998) Observer reliability and agreement. In: Armitag P, Colton T, editors. Encyclopedia Biostatistica Vol 4. 3123-3128. Chichester: John Wiley & Sons, Ltd
33. Streiner DL, Norman GR (2003) Health measurement scales: a practical guide to their development and use. New York: Oxford University Press
34. Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2005) Intra- and interrater reliability of the Ergo-Kit FCE method in adults without musculoskeletal complaints. Arch Phys Med Rehabil 86: 2354-2360
35. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG (2002) Interrater reliability in clinical trials of depressive disorders. Am J Psychiatry 159: 1598-1600
36. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF (2003) A comparison of two time intervals for test-retest reliability of health status instruments. J Clin Epidemiol 56: 730-735
37. Astrand P, Rodahl K, Dahl HA, Stromme SB (2003) Textbook of Work Physiology: Physiological Bases of Exercise. Champaign, IL: Human Kinetics
38. Wilmore JH, Costill DL(1999) Physiology of Sport and Exercise. Champaign, IL: Human Kinetics
39. Boadella JM, Sluiter JK, Frings-Dresen MHW (2003) Reliability of upper extremity tests measured by the Ergos Work Simulator: a pilot study. J Occup Rehabil 13: 219-232
40. Von Korff M, Ormel J, Keefe FJ, Dworkin SF (1992) Clinical section: Grading the severity of chronic pain. Pain 50: 133-149
41. Underwood MR, Barnett AG, Vickers MR (1999) Evaluation of two time-specific back pain outcome measures. Spine 24: 1104-1112
42. Smith BH, Penny KI, Purves AM, Munro C, Wilson B, Grimshaw J, Chambers WA, Smith WC (1997) The Chronic Pain Grade questionnaire: validation and reliability in postal research. Pain 71: 141-147
43. Vlaeyen JWS, Kole-Snijders AMJ, Rotteveel AM, Ruesink R, Heuts PHT (1995) The role of fear of movement/(re)injury in pain disability. J Occup Rehabil 5: 235-252
44. Roelofs J, Goubert L, Peters ML, Vlaeyen JW, Crombez G (2004) The Tampa Scale for Kinesiophobia: further examination of psychometric properties in patients with chronic low back pain and fibromyalgia. Eur J Pain 8: 495-502

45. Swinkels-Meewisse EJCM, Swinkels RAHM, Verbeek ALM, Vlaeyen JW, Oostendorp RA (2003) Psychometric properties of the Tampa Scale for kinesiophobia and the fear-avoidance beliefs questionnaire in acute low back pain. Manual Therapy 8: 29-36
46. Fleiss JL (1986) The Design and Analysis of Clinical Experiments. New York: Willey Classics Library
47. Tinsley HEA (1975) Interrater reliability and agreement of subjective judgements. J Couns Psychol 22: 358-376
48. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86: 420-428
49. Dimitrov D, Rumrill P, Fitzgerald S, Hennessey M (2001) Reliability in rehabilitation measurement. Work 16: 159-164
50. Hripcsak G, Heitjan DF (2002) Measuring agreement in medical informatics reliability studies. J Biomed Informs 35: 99-110
51. Essendrop M, Schibye B, Hansen K (2001) Reliability of isometric muscle strength tests for the trunk, hands and shoulders. Int J Indust Ergonomics 28: 379-387
52. Horneij E, Holmström E, Hemborg B, Isberg P, Ekdahl C (2002) Inter-rater reliability and between-days repeatability of eight physical performance tests. Adv Physiother 4: 146-160
53. Perry J, Weiss WB, Burnfield JM, Gronley JK (2004) The Supine Hip Extensor Manual Muscle Test: a reliability and validity study. Arch Phys Med Rehabil 85: 1345-1350
54. Roy MAG, Doherty TJ (2004) Reliability of hand-held dynamometry in assessment of knee extensor strength after hip fracture. Am J Phys Med Rehabil 83: 813-818
55. Lechner DE, Jackson JR, Roth DL, Straaton KV (1994) Reliability and validity of a newly developed test of physical work performance. J Occup Med 36: 997-1004
56. Saunders RL, Beissner KL, McManis BG (1997) Estimates of weight that subjects can lift frequently in functional capacity evaluations. Phys Ther 77: 1717-1728
57. Matheson LM, Mooney V, Grant JE, Affleck M, Hall H, Melles T, Lichter RL, McIntosh G (1995) A test to measure lift capacity of physical impaired adults. Part 1-Development and reliability testing. Spine 20: 2119-2129
58. Gross DP, Battie MC (2002) Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther 82: 364-371
59. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN (2003) Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic low back pain. J Occup Rehabil 13: 207-218

60. Reneman MF, Dijkstra PU, Westmaas M, Goeken LNH (2002) Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. J Occup Rehabil 12: 269-275
61. Isernhagen SJ, Hart DL, Matheson LM (1999) Reliability of independent observer judgments of level of lift effort in a kinesiophysical functional capacity evaluation. Work 12: 145-150
62. Reneman MF, Jaegers SMHJ, Westmaas M, Goeken LNH (2002) The reliability of determining effort level of lifting and carrying in a functional capacity evaluation. Work 18: 23-27
63. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1: 307-310
64. Beckerman H, Roebroeck ME, Becher JG, Bezemer PD, Verbeek AL (2001) Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 10: 571-578
65. Clark DJ, Condliffe EG, Patten C (2006) Reliability of concentric and eccentric torque during isokinetic knee extension in post-stroke hemiparesis. Clin Biomech (Bristol Avon) 21: 395-404
66. Gagnon D, Nadeau S, Gravel D, Robert J, Bélanger D, Hilsenrath M (2005) Reliability and validity of static knee strength measurements obtained with a chair-fixed dynamometer in subjects with hip or knee arthroplasty. Arch Phys Med Rehabil 86: 1998-2008

Chapter 5

The Utility of Functional Capacity Evaluation: the opinion of physicians and other experts in the field of Return to Work and Disability Claims

An adapted version of: Haije Wind, Vincent Gouttebarge, P.Paul F.M. Kuijer, Judith K. Sluiter, Monique H.W. Frings-Dresen. The utility of Functional Capacity Evaluation: the opinion of physicians and other experts in the field of return to work and disability claims. International Archives of Occupational and Environmental Health 2006; 79: 528-534

Abstract

Objectives: This expert poll explored how Dutch experts perceive the utility of FCE (Functional Capacity Evaluation) for return to work (RTW) and disability claim (DC) assessment purposes.

Methods: Twenty-one RTW case managers and 29 DC experts were interviewed by telephone using a semi-structured interview schedule.

Results: The RTW case managers valued the utility of FCE on a scale of 0-10. Their mean valuation was 6.5 (SD 1.5). The average valuation for DC experts was 4.8 (SD 2.2). Arguments in favor of FCE were (1) its ability to confirm own opinions and (2) the objectivity of its measurement method. Arguments against FCE were (1) the redundancy of the information it provides and (2) the lack of objectivity. Indications for FCE were musculoskeletal disorders, a positive patient self-perception of ability to work, and the presence of an actual job. Contraindications for FCE were medically unexplained disorders, a negative patient self-perception of ability to work, and the existence of disputes and legal procedures.

Conclusions: The responding RTW case managers perceived FCE to be more useful than the responding DC experts. The question of whether the arguments presented for and against the utility of FCE are valid is one that should be addressed in a future study.

5.1 Introduction

Assessment of functional physical capacity for work is a complex process. In the Netherlands, physicians who support disabled workers in their efforts to return to work, as well as physicians who value disability claims have few instruments to assess this capacity. Functional Capacity Evaluation is acknowledged as a potentially valuable tool for evaluating physical work capacity for the purpose of return to work and for the assessment of disability claims [1,2].

Functional Capacity Evaluation (FCE) methods aspire to offer systematic, comprehensive, and multi-faceted approaches designed to measure the current functional physical capacity of individuals with musculoskeletal complaints in relation to their work-related tasks [2-5,]. FCEs rely on a battery of standardized tests that reflect work-related activities, such as standing, walking, lifting, carrying and reaching [6-8]. FCE assessments evaluate performances of both short and long durations. During the tests, several factors are systematically reported to gain insight into the worker's functional physical abilities. These factors include the load lifted, working height, working distance, manipulation velocity, heart frequency, coordination, degree of pain and fatigue. FCE has been introduced and used for the assessment of work capacity and rehabilitation therapy in the USA, Canada, Australia and several European countries, such as Switzerland, Germany and the Netherlands [4,9-12]. In light of this, FCE is a potential instrument that physicians working in the field of return to work and those assessing disability claims could use to gain insight into functional physical capacity. Aside from questions about the FCE's reliability and validity, this instrument's utility can also be seen as an important issue.

The utility of an instrument is strongly related to the purposes for which it is used. According to Matheson and colleagues[13] utility refers, besides other aspects, to the suitability of the evaluation for the intended purpose, the extent to which it meets the needs of the client and the referrer. The utility of tests and assessments is determined by the extent to which the results facilitate planned interventions [14]. The concept of utility encompasses three distinct dimensions: utility on an individual level [15-19], utility on the organizational level [20], and utility of the instrument itself [21,22]. The last dimension of utility concerns the psychometric properties of an instrument. To the author's knowledge, the present study is the first to focus on the individual-level utility of FCE (or, more accurately, experts' perceptions of its individual-level utility) for the purposes of return to work and disability claims. Experts consider an instrument useful to the extent that it supplies new information or information that confirms that which was already was known.

Experts in the field of assisting disabled workers to return to work are occupational physicians. Other professionals involved in the process of enabling temporarily disabled subjects to return to work include rehabilitation physicians and reintegration advisors that work for reintegration organizations or municipalities, involved in reintegration. In this study, the latter experts are referred to as RTW (return to work) case managers. Experts in the field of assessment of disability claims are insurance physicians. Injury claim physicians and judges at administrative law courts are also involved in the process of assessing disability claims in the Netherlands. These experts are referred to as DC (disability claim) experts.

The present study describes the perceptions of the experts mentioned above concerning the utility of FCE for the purpose of supporting the process of return to work as well as for assessing disability claims and the determining conditions they keep in mind. Determining conditions are conditions that determine whether it is appropriate to perform an FCE assessment. These perceptions were gauged based on the following research questions:

- How do RTW case managers and DC experts perceive the utility of FCE for their work?
- What arguments do RTW case managers and DC experts present to describe the utility of FCE?
- What determining conditions do RTW case managers and DC experts consider with regard to the utility of FCE?

5.2 Method

Experts

A variety of procedures were used to contact experts for the study. In conducting an expert poll, it is important to try to contact as many potential experts as possible. For that reason, we carried out an extensive search for possible participants. Addresses of experts were retrieved through professional and branch organizations, the Internet, or through referrals from individual members of groups of experts. The experts were subsequently contacted by letter, telephone, or e-mail to determine whether they were familiar with FCE and, if so, to invite them to participate in the study. Familiarity with FCE was defined as experience with a request for an FCE assessment or experience using the outcome of an FCE test personally in one's work. Each expert who agreed to participate received a letter describing the aim of the research prior to the interview. Appointments were then made for interviews by telephone. All the respondents who were willing to participate and who met the inclusion criteria were included in the study. No other selection procedures were used.

Procedure and set-up

The authors formulated a list of questions for investigating the three research questions cited above. A single interviewer (HW) asked the questions by telephone, using a semi-structured interview schedule. Respondents were first asked to rate their perceptions of the utility of FCE on a scale of 0-10. Next, they were asked to state whether they considered FCE useful or not useful (yes or no), and how they had arrived at that judgment. Third, respondents were asked whether or not they found FCE useful as a prognostic instrument for future work ability and to explain their evaluations. Finally, the interviewer asked respondents to list any determining conditions that they considered applicable to the use of FCE as an assessment tool.

Each telephone interview was expected to last 20 minutes. The interviewer recorded responses to the three main questions on a scoring form, which contained the following answer categories:

I. Responses concerning the utility of FCE were classified into nine predefined categories, four of which represented positive evaluations and five represented negative evaluations (see Table 1).

II. Responses concerning the prognostic value of FCE assessment were classified into five categories, two of which represented positive evaluations and three represented negative evaluations. It was also possible for respondents to have no opinion on this issue (see Table 1).

III. Responses concerning the determining conditions were classified into six categories. Two of these represented the disorder and the assessment; one represented the patient and one the FCE instrument (Table 1). Respondents were asked to name any determining conditions that came to mind. They were then presented with a list of the different determining conditions and asked to state whether or not they considered each item on the list to be valid. Respondents indicating that an item was valid for FCE assessment were asked whether the condition would be an indication or contra-indication for an FCE assessment.

Table 1: Categories of arguments for the utility of FCE, the prognostic value of FCE and determining conditions for performing an FCE assessment

Arguments for the utility of FCE	FCE confirms the experts' own judgment FCE is objective FCE involves the observation of behavior and can predict behavior in the work environment FCE provides patients and others (e.g. employers and treating physicians) with insight into patients' ability to work	
Arguments against the utility of FCE	FCE provides no new information FCE is not objective because the patient has influence on the outcome FCE does not measure all aspects that are relevant for work FCE measures aspects of work that are irrelevant to DC assessments (e.g. fitness, constitution, or fatigue) FCE does not resemble actual working situations	
Arguments for the prognostic value of FCE	FCE shows opportunities for recovery FCE facilitates preventive recommendations concerning the tasks and content of suitable work	
Arguments against the prognostic value of FCE	FCE adds no complementary information FCE is a momentary assessment FCE is not a valid measure of functional capacity	
Determining conditions for performing an FCE assessment	Type of disorder: Severity of disorder: Self perception of patient: Quality of instrument: Moment of assessment: Context of assessment: 	musculoskeletal disorders syndromes of medically unexplained symptoms neurological and other disorders very serious disorders medical contra-indication for FCE positive towards ability to work negative towards ability to work good quality FCE instrument poor quality or lack of quality FCE instrument stable medical situation progressive disorder early in process of return to work rehabilitation in a deadlock actual availability of a workplace dispute situation and/or legal procedure

Data analysis

The answers from the scoring forms were coded and recorded in an electronic database. After the interviews with all of the experts were complete, a random sample of 10 scoring forms was coded by the second author (VG) and compared to the coding of the first author. Where

differences occurred between the first author's original coding and that of the second author, an extra scoring form was added to be checked by the second author.

For each group of experts, the mean scores, standard deviations and ranges were calculated to arrive at the overall utility score. Differences between expert groups concerning whether FCE was considered useful were tested, using a Mann-Whitney test. A p-value < 0.05 was considered statistically significant. A rank bi-serial correlation coefficient [23] was calculated to check on the consistency between the utility score and the utility valuation of FCE. A correlation coefficient < 0.51 was interpreted as poor; coefficients ≥ 0.51 and < 0.75 were considered moderate and those ≥ 0.75 were good [24]. For each expert group, the total number of arguments per category of FCE utility regarding RTW and DC was determined. The same procedure was followed for the categories regarding the utility of FCE as a prognostic instrument and the categories of determining conditions regarding the application of FCE.

5.3 Results

Of the potential pool of some 2100 experts that were asked to participate in this expert poll, 109 respondents (individuals and organizations) replied to the invitation. Of those 109 respondents, 82 subjects applied themselves, or were referred as possible participants in this study. A total of 50 subjects met the inclusion criteria; 25 were excluded, mainly because they had no experience with FCE. In seven cases, the experts could not be contacted or fell in another category of experts. This was especially the case with the occupational physicians, who appeared to be working as insurance physicians, or as RTW consultants. These participants were included in those experts groups. All the respondents who were willing to participate were included in the study. In other words, no selection procedure was used. Twenty-one of these experts were working as RTW case managers, and 29 were DC experts. Table 2 presents information concerning the selection of the eight groups of experts in this study.

Table 2: Expert groups classified according to work setting, retrieval point of addresses, total number of addresses contacted, method of contact, number of responses and number of entries.

Expert group	Retrieval of addresses	Number of addresses	Method of contact	Number of responses	Number of entries	Participants
RTW case managers						
Occupational physicians	NVAB*	2000	Letter; reply strip	25	25	8
Rehabilitation physicians	VRA*	9	e-mail; telephone	7	9	5
Consultants (RTW centers)	BOREA*	9	Telephone	9	10	4
Consultants (municipalities)	Four major cities and FCE providers	7	Telephone	7	7	4
DC experts						
Insurance physicians, public	UWV*	61	e-mail	40	16	15
Insurance physicians, private	GAV*	1	Telephone	7	7	7
Injury claim physicians	MAS*	7	Telephone	5	5	4
Administrative law judges	List of addresses	19	Letter	9	3	3

*NVAB: Dutch Professional Organization for Occupational Physicians; VRA: Dutch Professional Organization for Rehabilitation Physicians; BOREA: Branch Organization for Return to Work Centers; UWV: National Institute for Employee Benefits Schemes; GAV: Dutch Organization of Insurance Physicians at Private Insurance Companies; MAS: Medical Advisors for Injury Victims.

Eight occupational physicians and five rehabilitation physicians participated as experts in the area of managing the process that enables disabled workers to return to work. Other experts in this area included eight consultants: four from municipalities and four from organizations that manage the process in which disabled workers to return to work. The experts on assessing disability claims included: 15 insurance physicians in the public sector; seven insurance physicians in the private sector; four physicians working as insurance physicians for injury claims; and three magistrates at courts of administrative law. Table 3 presents information concerning the respondents and their work experience.

Table 3: Number of experts (N), mean range of work experience (in years), and mean score 0-10 (SD, range and p-value) for FCE utility, by expert group

Subjects	N	Experience (in years)		Utility of FCE		
		Mean	Range	Mean	SD	Range
Return to Work case managers	21	12.5	1-25	6.5 *	1.5	1-8
Disability claims experts	29	11.4	2-26	4.8	2.2	0-8
Total	50	11.0	1-26	5.5	1.5	0-8

SD: standard deviation; *: significant difference from disability claims experts (p = 0.001)

None of the telephone interviews exceeded the expected 20 minutes. Twelve forms were scored by the second author (VG).

Utility of FCE

Table 3 displays the mean FCE utility score for each expert group. Return to work experts assessed the utility of FCE on average at 6.5 (SD 1.5); disability claim experts averaged 4.8 (SD 2.2). RTW case managers perceived FCE to be significantly more useful than did DC experts ($p = 0.001$). Two-thirds of all participants considered FCE useful; however, 18 did not consider it so. Almost all of the RTW case managers considered FCE useful. Thirteen of the 29 DC experts considered it useful, and 16 did not consider it useful. The rank bi-serial correlation coefficient between the FCE utility score and the FCE utility valuation was 0.83, which, as explained earlier, was considered good.

Of the four categories of arguments for the utility of FCE, experts referred most often to confirmation of (their own) judgment and the objectivity of measurement. RTW Case managers also mentioned insight regarding patients' functional capacity for others as an argument supporting the utility of FCE (Table 4). Of the five categories of arguments against the utility of FCE, experts most frequently cited the instrument's inability to provide new information (redundancy of information) and the vulnerability of its objectivity to the effects of patient malingering. These arguments were particularly prominent among DC experts.

Table 4: Frequency with which arguments were mentioned for or against the utility of FCE, by category of expert.

	Return to work	Disability claim	Total
Arguments for FCE utility:	N = 19	N = 13	N= 32
Confirms own judgment	14/19	6/13	20/32
Objective measurement method	13/19	10/13	23/32
Observation of behavior	2/19	1/13	3/32
Insight into work ability for patient and others	11/19	3/13	14/32
Arguments against FCE utility:	N = 2	N = 16	N=18
Adds no new information	2/2	10/16	12/18
Is not objective, patient has too much influence	1/2	10/16	11/18
Does not measure all aspects/measures irrelevant aspects	0/2	6/16	6/18
Does not resemble actual work situation	0/2	2/16	2/18

N: number of subjects

Ten of the experts participating in this study considered FCE useful as a prognostic instrument, and 35 did not. Five respondents did not answer this question, as they had no knowledge of any studies that have been published in the scientific literature regarding the prognostic value of FCE. Thirty of the 35 participants who did not consider FCE to have any prognostic value supported their assessments with the observation that FCE represents only a momentary judgment of functional physical capacity.

Prerequisites

Nearly all of the respondents (90%) mentioned determining conditions, with contra-indications outnumbering indications in relation to the utility of an FCE assessment (Table 5). DC experts listed more determining conditions than did RTW case managers.

Musculoskeletal disorders were cited as the type of disorder for which an FCE assessment could be useful, while medically unexplained disorders were cited as contra-indicative. Negative perceptions by patients of their own ability to work were the most prominent indication cited by RTW case managers against performing such an assessment. Half of the DC experts cited the existence of disputes, legal procedures, and injury claim procedures as contra-indicative for the utility of FCE.

Table 5: Frequency with which determining conditions were mentioned for or against the utility of FCE, by category of expert.

		Return to Work	Disability Claims	Total
Indications for FCE assessment		N = 11	N = 18	N = 29
Type of disorder	Musculoskeletal disorders	4/11	5/18	9/29
	Other disorders	4/11	3/18	1/29
Patient self-perception	Positive attitude	0/11	5/18	5/29
Quality of FCE instrument	Good quality	2/11	5/18	8/29
Timing of FCE assessment	Early rehabilitation process	1/11	3/18	4/29
	Progress in rehabilitation process	2/11	3/18	5/29
Context of FCE assessment	Availability of a job	0/11	8/18	8/29
Contra-indications for FCE assessment		N = 13	N = 18	N = 31
Type of complaint	Unexplained medical symptoms	2/13	8/18	10/31
	Other disorders	3/13	3/18	6/31
Severity of disorder	Serious medical disorders	2 /13	2/18	4/31
	Medical contra-indication FCE	0/13	1/18	1/31
Patient self-perception	Negative attitude	9/13	13/18	22/31
Quality of FCE instrument	Low/lacking quality	0/13	2/18	2/31
Timing of FCE assessment	Medically unstable situation	3/13	3/18	6/31
Context of FCE assessment	Juridical procedure/dispute situation	-	9 /18	9/18*
	Injury claim procedure	-	8 /18	8/18*

N: number of subjects; * was only presented to DC experts

5.4 Discussion

This expert poll has focused on arguments from various experts concerning the utility of FCE. The expert respondents were identified through an exhaustive search for expertise relevant to the study. RTW case managers valued the utility of FCE for their work setting significantly higher than did DC experts. Arguments cited in favor of FCE were its potential for confirming one's own opinion and its objectivity of measurement. In contrast, arguments against using FCE were: the instrument's inability to offer new information and its subjectivity of measurement. Musculoskeletal disorders were considered the category of disorders for which FCE is useful. The patients' perceptions of their abilities to work and the context in which the FCE takes place were cited as the most important prerequisites for effective application of FCE assessment.

One of the most remarkable results is the discrepancy between the number of experts that were approached and the number that responded to the invitation. This appears to indicate that

FCE is not an instrument frequently used by the groups of experts approached. However, the aim was not to establish how familiar FCE is to these groups of experts, but to determine how they value its utility. Nonetheless, the small number of participants does limit the extent to which generalizations can be made about the value of FCE. Incidentally, these experts' lack of familiarity with this instrument was somewhat surprising, as Dutch occupational and insurance physicians have few instruments for assessing the functional physical capacity of workers with musculoskeletal disorders. In light of that, one would expect an instrument, developed to assess functional physical capacity in work situations to be more widely known, especially among these groups of experts.

The group of RTW case managers valued the utility of FCE higher than did the group of DC experts. This is a remarkable finding, though as mentioned above, a certain caution is necessary in drawing conclusions given the small number of respondents. A considerable percentage of the experts on assessing disability claims questioned the utility of FCE assessments, and appeared, therefore, to rely more on judgments based on their own assessments. This difference in opinion about the utility of FCE cannot be explained based on the results of this study. One could speculate that this difference in the valuation of the FCE's utility is attributable to the difference in context: DC experts assess disability claims, whereas RTW case managers try to facilitate return to work. This implicates that DC experts operate in a more legal context, and RTW case managers in a more practical context.

At the individual level, the utility of an assessment instrument hinges on one question: do the test results confirm what was already known, or do they provide new insight into what was unknown? The arguments for – as well as those against – the utility of FCE fit within this definition of utility. The argument regarding confirmation of one's own opinion fits this definition. Similarly, the argument regarding the objectivity of the measurement can also be seen as confirming, or completing the expert's opinion about a patient's functional physical capacity. The argument that an FCE assessment provides no new information can be directly related to the definition of utility. The fear that the subject could have too much influence on the outcome is an argument that applies to the confidence the expert can have in the information presented from an FCE assessment. Information that cannot be trusted cannot be useful.

Musculoskeletal disorders were mentioned as the category of disorders for which FCE could be a useful instrument. This falls in line with several studies on different aspects of the validity and reliability of FCE, where subjects with musculoskeletal disorders were tested [25-28]. Since FCE focuses on functional physical capacity, it measures physical abilities.

The context in which an FCE measurement is performed was cited as an important aspect of the instrument's utility. For instance, in the case of workers' compensation for claimants with work-related low back pain, such as those in Gross and colleagues' study [29], the FCE's utility is valued differently than it is in a context of assessing workers as candidates for a job [30]. Moreover, in the case of a dispute or legal procedure, an individual's willingness to cooperate in an FCE test can be subject to question. These aspects regarding the utility of an FCE assessment were cited by several experts. In other words, the sincerity of effort appears to be an issue that the experts bear in mind. Where there is no willingness to cooperate in an FCE test, the person's abilities will be misrepresented, and conclusions based on the assessment can be erroneous [31]. FCE methods were designed to include procedures to monitor the subject's efforts in performing the test. The purpose of this is to gain an impression of the sincerity of those efforts. Rudy and colleagues point out that, regardless of the FCE's primary objective, it is important to recognize that the instrument conducts behavioral assessments and that many environmental factors can influence or bias its results [32]. This falls in line with an argument that some participants cited to support their opinion that FCE is not useful, namely that FCE does not measure all the aspects important to work capacity and that aspects that should not play a role in disability claim assessments influence the outcome of an FCE assessment.

Is FCE considered a useful instrument for assessing functional physical capacity? It is not possible to answer this question conclusively based on the results of this study. The answers that were given consisted of arguments for and against use of the instrument. RTW case managers appeared to have positive views of the FCE's utility, unlike most of the DC experts, who were negative. Over the past few years, several studies have been published on the reliability and validity of FCE [24,33-35]. However, the question of whether FCE is useful to potential users has remained unanswered. Utility is an important aspect of any instrument, and FCE is certainly no exception. This study explored the opinion of experts. Another approach to studying the utility of FCE would be to examine the influence of FCE information on the judgment of an expert in an experimental study. The task of assessing the functional physical capacity of subjects with musculoskeletal disorders is complex. What is more, it usually involves subjective judgments, both on the part of the investigator and the person being investigated. An FCE assessment provides information on functional physical abilities, using a different approach, namely by measuring the performance of work-related activities. FCE assessments are developed as instruments that are intended to objectively measure functional

physical abilities in work-related activities, an objective that continues to be a major issue in society today.

Reference list

1. Lechner D, Roth D, Straaton K (1991) Functional capacity evaluation in work disability. Work 1: 37-47
2. Strong S (2002) Functional Capacity Evaluation: the good, the bad and the ugly Occup Ther 5-9
3. King PM (1998) Sourcebook of occupational rehabilitation. New York: Plenum Press
4. King PM, Tuckwell N, Barrett TE (1998). A critical review of functional capacity evaluations Phys Ther 78: 852-866
5. Vasudevan SV (1996) Role of functional capacity assessment in disability evaluation. J Back Musculoskelet Rehabil 6: 237-248
6. Harten JA (1998) Functional Capacity Evaluation. Occup Med-State Art 13: 209-212
7. Innes E, Straker L (2002) Workplace assessments and functional capacity evaluations: Current practices of therapists in Australia. Work 18: 51-66
8. Tuckwell NL, Straker L, Barrett TE (2002) Test-retest reliability on nine tasks of the Physical Work Performance Evaluation. Work 19: 243-253
9. Deen M, Gibson L, Strong J (2002) A survey of occupational therapy in Australian work practice. Work 19: 219-230
10. Innes E, Straker L (2003) Attributes of excellence in work-related assessments. Work 20: 63-67
11. Kaiser H, Kersing M, Schian H-M, Jacobs A, Kasprowski D (2000) Value of the Susan Isernhangen Evaluation of Functional Capacity Scale in medical and occupational rehabilitation [Der Stellenwert des EFL-Verfahrens nach Susan Isernhagen in der medizinischer und beruflichen Rehabilitation: in German] Rehabilitation 39: 297-306
12. Strong S, Baptiste S, Clarke J, Cole D, Costa M (2004) Use of functional capacity evaluations in workplaces and the compensation system: a report on workers' and report users' perceptions. Work 23(1): 67-77
13. Matheson LN, Mooney V, Grant JE, Leggett S, Kenny K (1996) Standardized evaluation of work capacity. J Back Musculoskelet Rehabil 6: 249-264
14. Telzrow CF, McNamara K (2001) New directions in assessment for students with disabilities. Work 17: 105-116
15. Aaron LA, Patterson DR, Finch CP, Carrougher GJ, Heimbach DM (2001) The utility of a burn specific measure of pain anxiety to prospectively predict pain and function: a comparative analysis. Burns 27: 329-334

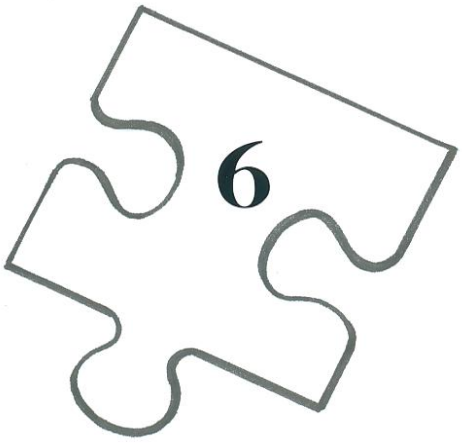
16. Duruoz MT, Cerrahoglu, Dincer-Turan Y, Kursat S (2003) Hand function assessment in patients receiving haemodialysis. Swiss Med Wkly 133 (31-32): 433-438
17. Feise RJ, Menke JM (2001) Functional Rating Index: a new valid and reliable instrument to measure the magnitude of clinical change in spinal conditions.. including commentary by Cherkin D. Spine 26(1): 78-87
18. Opasich C, Pinna GD, Mazza A, Febo O, Riccardi R, Riccardi PG, Capomolla S, Forni G, Cobelli F, Tavazzi L (2001) Six-minute walking in performance in patients with moderate-to-severe heart failure: Is it a useful indicator in clinical practice? Eur Heart J 22(6): 488-496
19. Simmonds MJ (2002) Physical function in patients with cancer: Psychometric characteristics and clinical usefulness of a physical performance test battery. J Pain Symptom Manage 24(4): 404-414
20. Van Dijk FJH, De Kort WLAM, Verbeek JHAM (1993) Quality assessment of occupational health services instruments. Occup Med 43 (suppl 1): S28-S33
21. Alderson M, McGall D (1999) The Alderson-McGall hand function questionnaire for patients with carpal tunnel syndrome: a pilot evaluation of future outcome measure. J Hand Ther 12(4): 313-322
22. Moore AD, Clarke AE, Danoff DS, Joseph L, Belisle P, Neville C, Fortin PR (1999) Can health utility measures be used in lupus research? A comparative validation and reliability study of 4 utility indices. J Rheumatol 26(6): 1285-1290
23. Portney LG, Watkins MP (2000) Part IV data analysis: correlation. In: Foundations of clinical research. Applications to practice. Upper Saddle River, New Jersey. Prentice Hall Health. 2nd (23): 503
24. Innes E, Straker L (1999) Reliability of work-related assessments. Work 13: 107-124
25. Abdel-Moty E, Fishbain DA, Khalil TM, Sadek S, Cutler R, Rosomoff RS, Rosomoff HL (1993) Functional capacity and residual functional capacity and their utility in measuring work capacity. Clin J Pain 9: 168-173
26. Gross DP, Battié MC (2002) Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther 82(4): 364-371
27. Saunders RL, Beissner KL, McManis BG (1997) Estimates of weight that subjects can lift frequently in functional capacity evaluation. Phys Ther 77(12): 1717-1728
28. Smith RL (1994) Therapists' ability to identify safe maximum lifting in low back patients during functional capacity evaluation. J Orthop Sports Phys Ther 18(5): 277-281

29. Gross DP, Battié, Cassidy JD (2004) The prognostic value of functional capacity evaluation in patients with chronic low back pain: part 1. Timely return to work. Spine 29:914-919 .. including comment by Oliveri M, Jansen T (2005). Spine 30: 1232-1234
30. Harbin G, Olson J (2005) Post-offer, pre-placement testing in industry. Am J Ind Med 47: 296-307
31. Simonsen JC (1996) Validation of sincerity of effort. J Back Musc Rehabil 6: 289-295
32. Rudy TE, Lieber SJ, Boston JR (1996) Functional capacity assessments: Influence of behavioral and environmental factors. J Back Musc Rehabil 6: 277-288
33. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN (2003) Test-retest reliability of the Isernhagen Work Sytems functional capacity evaluation in patients with chronic low back pain. J Occup Rehabil 12: 207-218
34. Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2005) Intra- and interrater reliability of the Ergo Kit® FCE methods in adults without musculoskeletal complaints. Arch Phys Med Rehabil 86: 2354-2360
35. Innes E, Straker L (1999) Validity of work-related assessments. Work 13: 125-15

Chapter 6

Effect of functional capacity evaluation information on the judgment of physicians about physical work ability in the context of disability claims

Haije Wind, Vincent Gouttebarge, P. Paul F.M. Kuijer, Judith K. Sluiter, Monique H.W. Frings- Dresen (Submitted)



Abstract

Objective

To test the influence of functional capacity evaluation (FCE) information on the judgment of the physical work ability of claimants with musculoskeletal disorders (MSD) by insurance physicians (IPs) in the context of disability claims.

Design

Pre/post-test controlled experiment within insurance physicians.

Setting

The assessment of work ability by IPs in the context of a statutory disability claim procedure. Two claimants from each IP participated. IPs scored the physical work ability of both claimants twice for 12 specified activities, using a visual analogue scale (VAS). In addition, one claimant underwent FCE while the other served as a control. The FCE information was added to the claimant's file. IPs then reassessed the physical work ability of both claimants based on the information collected since the initial assessment.

Participants

Twenty-seven IPs, and 54 claimants voluntary participated.

Main outcome measure

The difference between experimental and control groups is the number of shifts in physical work ability assessment measured by VAS scores.

Results

Receipt of FCE information caused IPs to change their assessment of the physical work ability of claimants with MSD on significant more activities than when they received no FCE information.

Conclusion

Provision of FCE information results in IPs to change their judgment of the physical work ability of claimants with MSD in the context of disability claim procedures. Change in judgment was in majority in line with the FCE results both in the direction of more as less physical work ability.

6.1 Introduction

The assessment of work ability in the context of long-term disability claim procedures is a complex matter, and the physicians who perform these assessments do not have many instruments to help them in this endeavour. Many people are subject to work-related illnesses or injuries, which may lead to long-term disability. In many countries, it is the statutory responsibility of physicians to assess the work ability of persons claiming disability benefit. It has been found that physicians are often unfamiliar with disability criteria and have little confidence in their ability to determine who is disabled and who is not [1]. The variability of impairment ratings among physicians is large and sometimes inconsistent with scientific evidence [2-4].

An important category of disorders presented to physicians in the context of assessing work ability for disability claims is that of musculoskeletal disorders (MSD). MSD is one of the major causes of disability, and the burden of MSD will increase in an ageing society [5]. The direct and indirect costs of chronic disability associated with these disorders in the USA and Canada is enormous [6].

There are only few instruments available to physicians engaged in the assessment of physical work ability that are both reliable and valid [7]. Some questionnaires have been found to have a high level of validity and reliability, but this is not the case for functional tests. Several studies on the reliability and validity of a number of functional tests, in particular functional capacity evaluation (FCE), have been performed of recent years [8-13]. FCE packages are batteries of tests designed to assess the physical ability of persons - especially (ex-)workers with MSD - to perform work-related activities [14]. The physical work capacity determined by FCE testing can be compared to the physical job requirements of the patient's occupation or to physical job requirements in general.

In the Netherlands, the ability of a patient to return to his or her old job or to undertake a new job is assessed by trained, certified insurance physicians (IPs) after 24 months of sick leave. IPs rely heavily on information received from claimants in such work ability assessments [15,16]. Assessing the physical work ability by IPs is like a diagnostic process, in which not the medical diagnose but the work ability is the target. As FCE information might be relevant for the judgment of the IP on the physical work ability, FCE could be added as an instrument in this process. The aim of the present study is, therefore, to explore the effect of information on the judgement of insurance physicians in the context of disability claim assessments of claimants with MSD. The research question is as follows:

- Does information derived from FCE tests lead insurance physicians to change their judgment of the physical work ability of claimants with musculoskeletal disorders?

6.2 Methods

A pre/post-test controlled experiment within subjects was used to answer the research question. To study the extent to which FCE information caused IPs to change their judgment of the physical work ability of a group of subjects with MSD in the context of a disability claims procedure, IPs assessed the work ability twice in an experimental group where the claimants underwent FCE tests after the first assessment, and in a control group where claimants did not undergo FCE tests. The medical Ethical Committee of the Academic Medical Center of the Universiteit van Amsterdam has approved this study.

Participants

Insurance physicians

In the Netherlands, statutory assessments of long-term disability claims are performed by IPs in the service of the Institute for Employee Benefit Schemes (UWV). The UWV is a semi-governmental organization that employs 566 IPs. One hundred IPs, selected at random, were invited to participate in the study. Fifty-four of these IPs complied with the inclusion criterion: they performed work ability assessments on long-term disability claimants, and were prepared to take part in the study. They all signed an informed consent form.

Claimants

Two claimants with MSD of each IP who were both seen in the context of a long-term disability claims procedure, were included in the study. Blinded for the IPs, the first signed an informed consent form and underwent FCE testing. A second claimant served as a control. The results of the FCE tests had no influence on the IP's statutory assessment of the claimant.

FCE test

The FCE test used in this study was the Ergo-Kit. This FCE relies on a battery of standardized tests reflecting work-related activities. A certified rater performed the 55 tests on each subject, following a standard protocol. The whole procedure took approximately three hours. If a medical contra-indication for FCE testing, e.g. heart failure or recent surgery, existed the claimant was excluded from the study. Reliability of Ergo-Kit lifting tests was found to be satisfactory in subjects with and without low back pain [8,9]. Content validity of the Ergo-Kit FCE is thought to be good, considering that the test procedures are fully described in a

manual, and that they are standardized, as well the procedure of drawing up a report. The tested activities are work-related and are, like the tested activities from other FCE assessment methods, derived from activities mentioned in the Dictionary of Occupational Titles (DOT)[15] .

Procedure

The work ability of each claimant was assessed by the IP in accordance with the statutory rules. IPs provided information about the study to claimants with MSD who were applying for a disability benefit or continuation of a disability benefit, and who complied with the inclusion criteria. The procedure is elucidated in Fig. 1.

The claimants were divided into two groups. Group 1 underwent FCE testing, while group 2 served as controls. As soon as an informed consent had been received from a patient in group 1, an appointment for FCE testing was made with the Ergo-Kit team. The FCE assessment always took place after the statutory assessment of the disability claim.

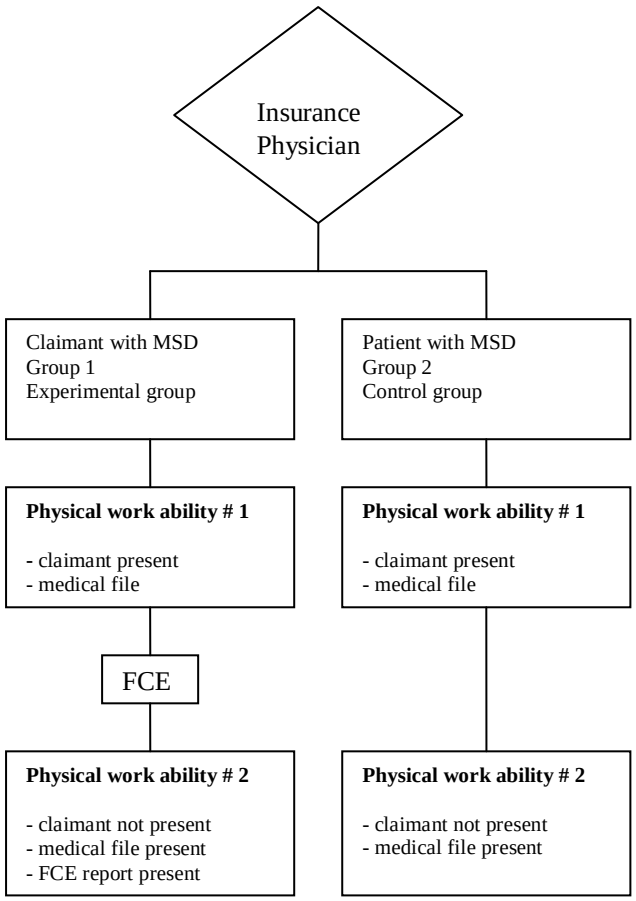


Fig. 1. Flow diagram of procedure used in the study

The claimants in the experimental group were tested in accordance with a standard Ergo-Kit protocol by 13 certified raters at 13 locations throughout the Netherlands. A report of the FCE tests performed was added to the claimant's file, and a copy was sent to the claimant.
The physical work ability of 54 claimants was scored twice by the 27 IPs in the context of long-term disability assessments. Half of this group underwent FCE tests, while the other 27 claimants formed the control group. The first claimant handled by a given IP who indicated willingness to participate in the study was assigned to the group that underwent FCE testing, without the knowledge of the IP. The second claimant of that IP was assigned to the group that underwent no FCE testing. In both cases, the IP assessed the work ability of each claimant twice: in the first group without (pre) and with (post) the information from the FCE assessment in connection with the information in the patient's file and, in the second group, based only on the information in the patient's file (pre-post). Claimants were always present during the first assessment, but not during the second; in the latter case, the IP reviewed the claimant's case on the basis of the information available in the file.

Outcomes

The characteristics of the IP, such as gender, age, years of experience with work ability assessment and familiarity with FCE were noted, as were the characteristics of the claimants, such as gender, age and location of disorder. The IPs were asked what information was used for the first and second assessment in both groups of claimants. The time interval between the IP's first assessment and the FCE test for each claimant was recorded.
Visual analogue scales (VAS) were used to record the results of the assessment of the physical work ability by the IP. The following twelve activities were rated: walking, sitting, standing, lifting or carrying, dynamic movements of the trunk, static bending of the trunk, reaching, movements above shoulder height, kneeling or crouching and three activities related to hand and finger movements (repetitive hand movements, specific hand movements and pinch or grip strength). These activities were selected from several questionnaires as being valid and useful for assessment of the physical work ability of subjects with MSD [7]. The VAS score ranged from 0 to 10 and was represented by a horizontal line, the length of 10 cm. The lower limit (0) was defined as complete lack of physical work ability for the activity in question compared to the situation before the claimant became disabled. The upper limit (10) was defined as no loss of physical work ability for that activity compared to the situation before onset of disability. The main outcome measure is a shift of more than 1 cm in the VAS score for work ability as determined for one of the twelve physical activities between the first

and second assessment carried out by each IP. A change of more than 1.0 cm between the two VAS scores for a given claimant was regarded as representing an intentional change in the IP's judgment of the physical work ability. This assumption was based on the outcome of an unpublished feasibility study. Six IPs assessed the physical work ability of claimants with MSDs in the context of disability claims and re-assessed the physical work ability after two weeks, based on the information in the claimants file. They scored the physical work ability by VAS scores for the same 12 activities as used in the present study. The shift between the first and second judgment was on an average of 0.66 cm (SD 0.5). Therefore, a shift of less than 1 cm is regarded as not intentional (average + 1 SD) and thus, not clinically relevant. Moreover, in a previous study in which VAS scales were used, a shift above 9 mm was considered to be clinically relevant [17].

Data analysis

The age of the IPs and of the claimants in the two groups, and the number of years' experience the IPs had in work ability assessment, were given as a mean value with the standard deviation. Other characteristics were noted as numbers and percentages.

A shift of more than 1 cm in the judgment of the IPs was considered a difference between first and second assessment. The McNemar Chi-square test for paired samples was used to test the significance of the effect of FCE information on IPs' judgement of physical work ability [18]. Tests were performed for the activities as a whole, as well as for the separate activities. The Bonferroni correction was applied, as a result of which a p-value smaller than 0.02 was considered to be statistically significant.

The relation between the results of the FCE assessment and the shift in judgment of the IPs was studied by classifying the results of the former for each activity into four separate classes. These classes were: 0-33% (class 1), 34-50% (class 2), 51-66% (class 3), 67-100% (class 4). These categories represent the ability to perform that activity during a whole day (higher number means better abilities). In addition, some activities, like kneeling, movements above shoulder height, dynamic movements of the trunk, and reaching, cannot be performed during the whole day according to the Ergo-Kit FCE. The maximum ability for these activities is set at 66% for the whole day. For these the classes were recalculated starting from 0% to 66% into four classes. Lifting, and grip and pinch force are presented in the FCE report in kilograms and interpreted by the test leader. The outcome and classes were: not possible, very low (class 1), low (class 2), average (class 3), high and very high (class 4). The outcome of eleven of the 12 activities (static bend work postures is not summarized in the

FCE report) was compared to the first VAS score by the IP. To this end, the VAS list was divided proportionally into four categories as in the FCE classification. The categories were: 0- 3.3 cm (class 1), 3.4-5.0 cm (class 2), 5.1- 6.6 cm (class 3), 6.7-10 cm (class 4). The classification for each activity in the four classes based on the first VAS score of the IP and the FCE result were compared. When they were similar, the expectation was that the IP would not alter his score on the second VAS scale during the second judgment. In the case of the FCE result showing either a lower or a higher class than the IP judgment, the expectation was that the IP would lower or raise his score on the VAS scale for that activity during the second judgment (a shift of more than 1 cm). The judgment was noted as 'corresponding' in the cases of no discrepancy between the first VAS score and FCE result, and when a lower FCE classification was followed by a lower classification by the IP on the second VAS score. Likewise, when the FCE classification was higher and the IP followed this classification by a raised judgement on the second VAS score, this was noted as 'corresponding'. Total numbers of corresponding outcomes were calculated, divided in unchanged outcome, lowered and raised outcome. All other cases were numbered as 'not-corresponding'. For these 'not-corresponding' outcomes, also the direction of the difference between the expected second VAS score and the actual second VAS score were noted.

It was possible to compare a total number of 297 activities by using this method. The scoring and analysis were performed independently by the first two authors (HW and VG). Any disagreements that remained after discussion, were resolved by consulting a third researcher.

The statistical analyses were carried out using SPSS version 13.

6.3 Results

Fifty-four IPs were willing to participate in the study and signed an informed consent form, a response rate of 54%. Mean age ± standard deviation (SD) of the IPs was 47 ± 7.1 years, and 56% of the IPs were male. They had 15 ± 7 years of experience in work ability assessment. Fifteen of the IPs were familiar with FCE. From 27 IPs claimants entered the study. From the other 27 IPs no claimants were included. These two groups of IPs did not significantly differ from each other in age, gender, and work experience. Only the Chi square test for familiarity with FCE of the IP and the participation of claimants from that IP in the study showed a significant difference, viz. that claimants from IPs who were, preceding the study, familiar with FCE participated more often than claimants from IPs who were not familiar with FCE. Fifty-four claimants (27 pairs from 27 IPs) indicated their willingness to participate in the study and signed an informed consent form during the study period, which extended from

November 2005 to February 2007. The mean time between the disability claim assessment and the FCE tests in the experimental group was 45 days (SD 24) The mean time between the first disability claim assessment and the re-assessment in the experimental group was 103 days (SD 43, range 39-184 days) and in the control group 106 days (SD 99, range 16-339 days). The high SD in the latter group is primarily caused by five exceptional long time intervals (more than 184 days). The characteristics of the claimants are described in Table 1. Between the claimants in the experimental group and the control group existed no statistical differences on age, gender and the location of disorders.

Table 1: Characteristics of claimants in experimental and control groups: gender, age, and location of disorder, together with number of other sources of information used in second assessment

	Experimental group (N=27)	Control group (N=27)
Male (number; percentage)	11 (41)	10 (37)
Female (number; percentage)	16 (59)	17 (63)
Age in years (mean, sd)	46 (.9)	43 (1.6)
Location of disorder: Upper extremity (No.,%)	3 (11)	1 (4)
Lower extremity(No.,%)	2 (7)	8 (30)
Back and neck (No.,%)	15 (52)	9 (33)
Combination (No.,%)	8 (30)	9 (33)
Other info sources for second assessment (No.)	0	2

Whether or not the provision of FCE information caused IPs to change their judgment of the physical work ability of claimants for the 12 specified activities by at least 1 cm on the VAS scale is presented in Table 2. In this table, the number of changed and unchanged activities both in the experimental and in the control group are presented. On single activities, there is no significant difference between the two groups. However, the provision of FCE information caused IPs to change their judgment of the physical work ability of claimants for the totality of 12 activities significantly more often than in the control group (p = .01).
The mean number of activities for which IPs changed their judgment to the above-mentioned extent in the experimental group was 4.7 (SD 2), compared with 4.0 (SD 2) in the control group. In the experimental group 52% of the number of activities remained unchanged, for 21% of the activities the judgment about work ability was lowered and for 27% of the

activities the judgment was raised. In the control group 63% of the number of activities remained unchanged, 22% was lowered and 15% was raised.

Table 2: Number out of 27 insurance physicians with changed or unchanged judgment of more than 1 cm on the VAS-scale for each activity during the second judgment compared to the first judgment in the experimental and control group (numbers), McNemar χ^2 test

	Experimental Group		Control Group		
	Changed	Unchanged	Changed	Unchanged	McNemar χ^2 test
Total of activities	155	169	124	200	0.01*
Walking	15	12	15	12	1.00
Sitting	9	18	12	15	0.51
Standing	15	12	12	15	0.61
Lifting/ carrying	15	12	13	14	0.77
Dynamic moving trunk	16	11	13	14	0.61
Static bending trunk	17	10	11	16	0.18
Reaching	12	15	8	19	0.39
Moving above shoulder height	16	11	9	18	0.09
Kneeling/ crouching	14	13	14	13	1.00
Repetitive movements hands	7	20	9	18	0.77
Specific movements hands	9	18	3	24	0.07
Pinch/ grip strength	10	17	5	22	0.13

* *p*- value: < .02

The two researchers agreed for 98% on the scoring and analysis of the comparison between the results of the second VAS score to the results in the FCE report and the first VAS score. Differences were not structural. Consensus was reached on the discussion points. Comparing these results, the conclusion is that the second VAS scores were in majority in accordance with the results of the FCE assessment. In 184 of the total of 297 times the IPs scored consistent with the expectation based on the FCE result. Of the number of 184, the IP's judgment and the FCE result were in line for 94 activities and therefore, no change took place. For 58 activities, the IPs lowered their judgment of work ability in line with the FCE result that showed that the patient performed lower than the IP had judged at the first assessment. For 32 activities, the IPs raised their judgment of work ability in line with the FCE result that showed higher results than rated at the first judgment. The judgment about dynamic movements of the trunk, kneeling, and pinch/ grip strength was most frequently lowered in

line with the FCE results. For 113 activities, the IPs did not follow the outcome of the FCE assessment and maintained in 67 cases their judgment despite the FCE result. In 22 resp. 24 cases, the IP lowered and raised the work ability for that activity in contrast to the outcome of the FCE assessment. The activity pinch/ grip strength showed the most difference between expected second VAS scores and FCE results. Reaching, movements above shoulder height, and pinch/ grip strength were the activities for which the IPs most often lowered their judgment in contrast to the FCE result. Six of the 27 IPs were responsible for 32% of the inequality between the second VAS score and FCE result, which means that an ample majority of IP judgments is in accordance with the FCE results.

6.4 Discussion

This study, based on a pre-post experimental design, evaluated the effect of FCE information on IPs' judgment of the physical work ability of disability benefit claimants with MSD. For the totality of activities the FCE information lead to a significant shift in work-disability claim assessment. Besides, for 11 of the 12 activities the judgment of the IPs is for 64% of the activities in line with the FCE report.

The first aspect to consider is whether the VAS system is a suitable means of recording physical work ability assessments made by IPs. Many studies have shown that VAS scales are indeed a reliable means of representing judgments [19,20]. It is the statutory duty of the IP to consider all the available information about the claimant's medical situation and ability to perform various tasks, and to decide on this basis whether he or she is fit to work, or is fully or partially disabled. There is no objective criterion that indicates whether this judgment is accurate. One argument in favour of the use of the VAS scale is that it may be more sensitive to changes in assessment than the Functional Ability List (FAL) the instrument currently used routinely by IPs for recording physical work ability in the context of disability. The FAL rates physical work ability on an ordinal scale in 2, 3, or 4 categories, and will probably not reflect relatively small changes.

The next main topic for consideration is the suitability of FCE as a source of supplementary information in work ability assessment. While suggestions have been made previously to include FCE information in the disability screening process, we believe that the present study is the first one actually to measure the influence of this information on the judgment of IPs in a claim procedure [21,22]. The study of Oesch et al. should be mentioned in this context [23]. The setting of the study was the assessment of work capacity for decisions

about medical fitness for work. The use of FCEs in that study improved the quality of medical Fitness for Work Certificates after rehabilitation. The focus on a rehabilitation intervention is the main difference with the present study in which the assessment of physical work ability is the main outcome and not the evaluation of a rehabilitation program. The similarity between both studies is the influence of FCE information on the judgment of IPs of work ability. This study was designed to allow the effect of FCE information on IPs' judgment of physical work ability to be studied in its natural setting – with the proviso that, in contrast to normal diagnostic routine, the IPs taking part in the present study could not refer claimants for FCE testing themselves. They were unaware whether claimants were participating in the study during the first work ability assessment. In the experimental group, the FCE report was the only new information added to the claimant's file during the review of the physical work ability. It so happened that new information from other sources was added to the files of claimants in the control group in two cases; this may also have influenced the IPs'judgment in these cases. No specific direction was found for the change in judgment between the initial assessment and the review: for some activities, the assessment tended to change from a higher to a lower ability, while for other activities the change tended to be in the reverse direction. This contrasts with the findings obtained in the study of Brouwer et al. [24], stating that the FCE result showed a higher level of physical work ability of patients with low back pain compared to the IP judgment. The difference between the study of Brouwer et al. and the present study is that the IPs in the Brouwer study at their judgment did not receive the FCE report contrasting to this study in which the IPs received the FCE report and were asked to reconsider their judgment using the FCE information. Besides, the patients in the study of Brouwer et al were in a rehabilitation program and not in a statutory disability claim assessment which makes the context of that study different from the present one.

The majority of judgments of IPs about the activities (64%) was in accordance with the FCE results. Because in half of these cases the result of the first IP judgment appearing in the first VAS score was in accordance with the FCE result, it could be expected that the second VAS score would likewise be in accordance with both FCE result and first VAS score. However, in the other 50% the FCE result was not in accordance with the first VAS score. IPs altered subsequently their judgment in the direction of the FCE results. The direction of the alteration was more often towards less work ability than towards more work ability. When there was an inequality between the judgment of the IP and the results in the FCE report, most frequently IPs did not alter their judgments. As stated before only a small part of the IPs is responsible for a large proportion of the discrepancies between FCE report outcomes and IP

judgments. This finding might justify the conclusion that the majority of IPs in this study is susceptible to FCE information and only a few IPs hold on to their own judgment, despite the contrary FCE information.

Concerning the difference in number of changes between the control and experimental groups, the explanation could also be an dissimilarity between the two claimant groups. While the experimental group had appreciably fewer disorders of the lower extremities, the disorders at the other locations were fairly evenly divided. In the experimental group, disorders of the back and neck and combined disorders occurred the most frequently. Disorders of the lower back and combined disorders both involve several different physical activities, which may explain why a wide-spectrum set of tests like FCE provides information that can lead IPs to change their judgment on a range of different activities. Although there seems to be an inequality regarding the location of disorders in the two groups, the size of it was not such that it has led to statistical differences and therefore, dissimilarity between the two claimant groups cannot be explained by this difference.

The time between the initial assessment of physical work ability by the IP and the FCE tests (45 days on average) determines the period between the two assessments carried out by the IP on each claimant. In our opinion, this relatively long time gap does not invalidate the results of the study. The claimants who undergo the assessments have been disabled for a long time. The initial assessment takes place after two years of sick leave – and even longer in the case of those claimants who come for re-assessment after having received disability benefit for some time. It seems implausible that their physical work ability will change considerably between the initial assessment and the FCE tests. In fact, the long period between the two has the advantage that during the FCE tests the claimant has no recollection of the initial assessment by the IP. The period between initial assessment and review by the IP is of less importance both in the experimental and control group, because the review is based solely on inspection of the claimant's file without any actual physical examination of the claimant.

The assessment of physical work ability in the context of disability claim procedures is a complex process, characterized by considerable uncertainty about the accuracy of the outcome and hence leaving ample room for changes in judgment. Information derived from FCE tests is of a different nature than the other information that IPs use in assessing the physical work ability of workers with MSD in disability claim procedures, which is largely anecdotal and provided by the claimant himself – or herself. The advantage of FCE information might be that it is performance-based.

This study shows that the provision of FCE information caused IPs to change their judgment of the physical work ability of disability claimants with MSD. Physical work ability is important in situations of disability claim procedures, like in this study, but also in RTW and rehabilitation programs. In all these procedures, return to work of the disabled worker is the main goal. Although the context of this study is related to procedures that are specific for this social security system, return to work of disabled workers is a world-wide issue. What is found in this study is that professionals do take information from an FCE assessment seriously enough to alter their judgement about the physical work ability in disability claim assessments of workers with MSDs. This finding supports the complementary value of FCE information in the assessment of disability claimants with MSD.

6.5 Conclusions

The research question was whether information of an FCE assessment has an effect on the IP judgment of claimants with MSDs in the context of disability claims. The results of the study indicate that inded there is, in the sense that IPs follow in majority the results of the FCE assessment and change their judgment about the physical work ability of the claimants that underwent an FCE assessment significantly more often than in the controlled situation. Therefore, FCE would seem to be a valuable new addition to IPs arsenal of instruments to support them in judging the physical work ability of claimants. The complementary value might be even higher when IPs can decide themselves whether or not to refer claimants for FCE testing

Reference list

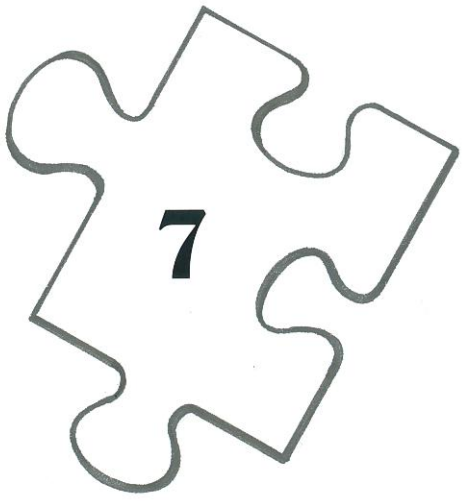
1. Zinn W, Furutani N (1996) Physician perspectives on the ethical aspects of disability determination. J Gen Intern Med. 11: 525-532
2. Patel B, Buschbacher R, Crawford J (2003) National variability in permanent partial impairment ratings. Am J Phys Med Rehabil. 82: 302-306
3. Carey TS, Hadler NM, Gillings D, Stinnett S, Wallstein T (1988) Medical disability assessment of the back pain patient for the social security administration: the weighting of presenting clinical features. J Clin Epidemiol. 417: 691-697
4. Rainville J, Pransky G, Indahl A, Mayer EK (2005) The physician as disability advisor for patients with musculoskeletal complaints. Spine. 30: 2597-2584
5. Brooks PM (2006) The burden of musculoskeletal disease – a global perspective. Clin Rheumatol. 25: 778-781
6. Baldwin M (2004) Reducing the costs of work-related musculoskeletal disorders: targeting strategies to chronic disability cases. J Electromyogr Kinesiol. 14: 33-41
7. Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2005) Assessment of functional capacity of the musculoskeletal system in the context of work, daily living, and sport: a systematic review. J Occup Rehabil. 15: 253-272
8. Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2005) Intra- and interrater reliability of the Ergo-Kit Functional Capacity Evaluation method in adults without musculoskeletal complaints. Arch Phys Med Rehabil. 86: 2354-2360
9. Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2006) Reliability and agreement of 5 Ergo-Kit Functional Capacity Evaluation lifting tests in subjects with low back pain. Arch Phys Med Rehabil. 87: 1365-1370
10. Reneman MF, Dijkstra PU, Westmaas M, Goëken LNH (2002) Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. J Occup Rehabil.12: 269-276
11. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goëken LNH (2003) Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic low back pain. J Occup Rehabil.13: 207-218
12. Gross DP, Battié MC (2002) Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther. 82: 364-371
13. Gross DP, Battié MC (2003) The construct validity of a kinesiophysical functional capacity evaluation administered within a workers' compensation environment. J Occup Rehabil. 12: 287-295

14. Hart DL, Isernhagen SJ, Matheson LN (1993) Guidelines for functional capacity evaluation of people with medical conditions. J Orthop Sports Phys Ther. 18: 682-686
15. Bont de A, Brink van den JC, Berendsen L, Boonk M (2002) Limited control of information for work disability evaluation. [De beperkte controle van de informatie voor de arbeidsongeschiktheidsbeoordeling: in Dutch] Ned Tijdschr Geneesk 146: 27-30
16. Knepper S (2002) Significance of medical data in work disability evaluation [De betekenis van medische gegevens bij de beoordeling van arbeidsongeschiktheid: in Dutch] Ned Tijdschr Geneesk 146: 6-8
17. Kelly AM. (1998) Does the clinically significant difference in visual analog scale pain scores vary with gender, age, or cause of pain? Am Emerg Med. 5: 1086-1089
18. Altman DG (1991) Comparing groups – categorical data. In: Altman: Practical statistics for medical research. Boca Raton, London, New York, Washington DC: Chapman and Hall pp 229-272
19. Zanoli G, Stromqvist B, Jonsson B (2001) Visual analog scales for interpretation of back and leg pain intensity in patients operated for degenerative lumbar spine disorders. Spine 26: 2375-2380
20. Anagnostis C, Mayer TG, Gatchel RJ, Proctor TJ (2003) The Million Visual Analog Scale: its utility for predicting tertiary rehabilitation outcomes. Spine 28: 1051-1060
21. Lyth JR (2001) Disability management and functional capacity evaluation: a dynamic resource. Work 16: 13-22
22. Liang MH, Dalroy LH, Larson MG, Partridge AJ, Abeles M, Taylor C, Fossel AH (1991) Evaluation of social security disability in claimants with rheumatic diseases. Ann Intern Med. 115: 26-31
23. Oesch PR, Kool JP, Bachmann S, Devereux J (2006) The influence of a Functional Capacity Evaluation on fitness for work certificates in patients with non-specific chronic low back pain. Work 26: 259-271.
24. Brouwer S, Dijkstra PU, Stewart RE, Goëken LNH, Groothoff JW, Geertzen JH. (2005) Comparing self-report, clinical examination and functional testing in the assessment of work-related limitations in patients with chronic low back pain. Disabil Rehabil 27: 999-1005

Chapter 7

Complementary value of functional capacity evaluation for physicians in assessing the physical work ability of workers with musculoskeletal disorders

Haije Wind, Vincent Gouttebarge, P. Paul F.M. Kuijer, Judith K. Sluiter, Monique H.W. Frings- Dresen. (Revised version of a submitted article)



Abstract

Objective

To study the complementary value of information from functional capacity evaluation (FCE) for insurance physicians (IPs) who assess the physical work ability of claimants with long-term musculoskeletal disorders (MSD).

Method

A post-test only design was used in the context of disability claims. Twenty-eight IPs participated in the study. Claimants with MSD formed the patient population. For each IP, the first claimant who agreed to participate was included in the study. The claimants underwent FCE in addition to the regular disability claim assessment. A self-formulated questionnaire was presented to the IPs after they viewed the FCE report. IPs were asked whether they perceived FCE information to be of complementary value to their judgment of the claimant's physical work ability investigated. We considered FCE information to be of complementary value if more than 66% of the IPs indicated as such. IPs were also asked whether FCE information led them to change their initial judgment about the claimant's physical work ability, and whether they felt this information made them more confident about their ultimate judgement. Finally, they were asked whether they planned to include FCE information in future disability claims and for what type of claimants. Differences between IPs who did or did not experience complementary value were explored.

Results

Nineteen (nearly 68% percent) of the IPs considered FCE information to be of complementary value for their assessment of claimants with MSD. Half of the IPs stated that the FCE information reinforced their judgment. All but four IPs changed their assessment after reading the FCE report. Sixteen IPs intended to involve FCE information in future disability claim assessments. There were no observed differences between the IPs who did or did not consider the FCE information to be of complementary value.

Conclusion

FCE information was found to have complementary value in the assessment of the physical work ability of claimants with MSD at present and in the future in the IPs opinion. Half of the IPs felt that this information reinforces their judgment in this context.

7.1 Introduction

Having work and being able to work are considered to be important requirements for being a full member of society. Work is an essential part of life for most of us. Inability to work, either because of unemployment, sickness or disability, has a negative impact on our quality of life [1]. Interventions aimed at assisting people in getting back to work should thus be encouraged. The assessment of the ability to work can play an important role in this context by permitting differentiation between those who can work and those who cannot. The former can be helped to return to work, while the latter are entitled to a temporary or permanent disability pension. The assessment of work ability can thus have a major impact both on the individual and on society as a whole.

In the Netherlands, insurance physicians (IPs) receive a four-year training in the assessment of work ability in persons who claim a disability pension after two years of sick leave. However, proper instruments for such assessment are lacking. The main source of information about the work ability of a claimant is the claimant him- or herself [2]. Since the claimant's opinion can differ considerably from that of the IP [3], there is a need for additional information (e.g. from physical examination or from the claimant's own doctor or specialist) if the work ability is to be reliably assessed. Only a few instruments are available for assessing the physical work ability of claimants with musculoskeletal disorder (MSD), and even these are only applicable only to special groups of claimants [4]. MSD is an important category of disorders in the context of disability claim assessments. In the Netherlands, about 30% of all disorders that led to disability claim assessments in 2004 involved the musculoskeletal system [5]. Musculoskeletal pain and its consequences are very common in the Dutch population 25 years and older [6]. MSD is also an important cause of absenteeism and disability in other European countries and the USA, leading to a high national illness burden [7,8].

Assessment of the physical work ability is a common practice in disability claim procedures. One instrument that might help IPs to assess the physical work ability of claimants with MSD is functional capacity evaluation (FCE). This approach makes use of highly structured, scientifically developed, individualized work simulators, designed to provide a profile of an individual's work-related physical and functional capabilities [9]. According to Harten [10], FCE offers a comprehensive, objective test that measures the individual's current functional status and ability to meet the physical demands of a current or prospective job. In particular, FCE provides information on physical work ability, being especially important in the assessment of disability in claimants with MSD and pain

syndromes [11]. In a previous study, we found that IPs who assess claimants with long-term disability have mixed opinions on the utility of FCE [12]. In fact, it appeared that only few physicians were familiar with FCE. Therefore, the topic of this study is whether FCE information can be of assistance to IPs in the assessment of the physical work ability of claimants, irrespective of their previous familiarity with the technique. This is a first step in the process of possibly introducing FCE in the process of assessing disability claims of claimants with MSDs. More specifically, the questions to which an answer was sought are:

- Is information derived from FCE of complementary value for an IP in the assessment of the physical work ability of claimants with MSD?
- Are there differences between IPs who do or do not consider the FCE information of complementary value in terms of personal characteristics of the IPs, themselves, or their claimants?
- Does FCE information lead IPs to change their assessment of a claimant's physical work ability, and/ or does it reinforce their judgment, both in the whole group and in the subgroups of IPs who do and do not consider FCE information of value?
- After having been introduced to FCE, are IPs likely to make use of FCE information in the assessment of claimants in the future? If so, for what groups of claimants? Also, what differences exist between the groups of IPs who do versus do not consider the FCE information to be of complementary value for future use?

7.2 Methods

The present investigation was designed as a post-test only study.

Participants

Insurance physicians

A total of 100 IPs who assess claimants for long-term disability benefits were randomly selected from a pool of 566 IPs who work for the Institute for Employee Benefit Schemes (UWV) in the Netherlands. This semi-governmental organization employs all IPs who perform statutory assessments of claimants for long-term disability benefit in the Netherlands. To test the hypothesis that 66% of the IPs conclude that FCE information has a complementary value for the assessment of physical work ability, under the assumption of the H_0 hypothesis of 40% [12], 28 IPs had to be included ($\alpha = 0.05$, $\beta = 0.8$). All participating IPs signed an informed consent form.

Claimants

Each IP gave information about the study to a number of MSD claimants who were due to be assessed in the context of long-term disability benefit claims. The information packet included an application form that the claimant could fill out and send directly to the researchers. The claimants could also indicate that they did not wish to participate and explain why (though they were not obliged to give any reason). The first claimant seen by a given IP who agreed to take part in the study underwent an FCE assessment after signing an informed consent form. The claimant received a copy of the FCE report. The Medical Ethical Committee of the Academic Medical Center, Amsterdam, approved the study. The study period was from November 2005 to February 2007.

Procedure

Each IP was asked to assess the physical work ability in accordance with the statutory rules for the claimant who had volunteered to participate in the study. After receiving the report of the FCE assessment from the FCE provider, this report was presented to the IP in combination with his own report in the patient's file. After reading the FCE report, the IP was requested to fill in a questionnaire in which he gave his opinion of the complementary value of the FCE information and stated whether the information led him to change his initial assessment. The statutory assessment of the claimant for the purposes of the disability benefit claim was based on the IP's initial judgement, i.e. the FCE information had no influence on this statutory assessment.

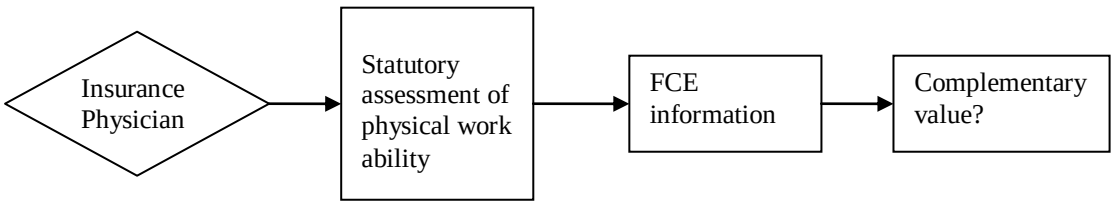


Fig 1: A flow diagram of the study design.

FCE test

The FCE instrument used in this study was the Ergo-Kit. This is comprised of a battery of standardized tests that reflect work-related activities. The standard protocol, containing 55 tests, was performed by certified raters and took approximately three hours to complete. The Ergo-Kit FCE was found to be reliable in subjects both with and without musculoskeletal complaints with respect to the lifting tests [13,14]. As a part of the test, claimants also filled in

the Revised Oswestry Pain questionnaire [15]. Claimants with a medical contra-indication for FCE, e.g. recent myocardial infarct, heart failure or recent surgery, were excluded from the test.

Outcomes

The questionnaire presented to all IPs contained three questions:

1) The IP was asked whether the FCE assessment had complementary value for the assessment of the physical work ability of the patient. The response choices were dichotomous: yes or no.

2a) For each of twelve activities selected on the basis of a previous study [4] as representative of the physical work ability of claimants with MSD (walking, sitting, standing, lifting/carrying, dynamic movement of the trunk, static bending of the trunk, reaching, movement above shoulder height, kneeling/crouching and three activities related to hand and finger movements), the IP was asked whether the FCE information caused him to revise his initial assessment of the claimant's ability upwards or downwards, or if it did not change the original assessment.

2b) The IP was asked whether the FCE information had reinforced his initial assessment of the claimant's physical work ability. The response categories were, again, dichotomous: yes or no.

3) Finally, the IP was asked whether he would consider using FCE in the future to support assessment of the physical work ability of disability benefit claimants; and if so, why, and for what groups of claimants in particular. If he did not favour the use of the FCE, the IP could also state their reasons for this view.

Data analysis

Descriptions of IPs and claimants were calculated. Age and years of experience of IPs were expressed as mean and standard deviation. The other characteristics of IPs, such as gender and familiarity with FCE, were noted in numbers and percentages. The age of the claimants was expressed as mean and standard deviation. The distribution of the location of the MSD (upper extremity, lower extremity, back and neck, or more than one location) was noted using numbers and percentages.

The answer to the first question in the IP questionnaire (whether FCE information was regarded as having complementary value for the assessment of physical work ability) was scored as affirmative when at least 66% of the IPs answered yes to this question. With regard

to the sub-question, characteristics of IPs and claimants that were believed to influence the answer of IPs about the complementary value of FCE information were classified. The characteristics selected for the IP group were work experience and familiarity with FCE. Work experience was found to be a factor that influences the way IPs come to their judgment about work ability [16,17]. Familiarity with FCE was judged to be another reason why IPs might think differently about the complementary value. It was deemed possible be that earlier contact with FCE information led to a negative opinion, as shown in the study about the utility of FCE information [12]. The characteristics registered in the claimant group were the location of the disorder and their working situation. Location of disorders could be a factor for differences in judgment of the complementary value of FCE information. Its possible that FCE information could be judged as more valuable to assessments of claimants with general disorders than specifically localized disorders. Work status is another characteristic of the claimants that could lead to a difference between the group of IPs that considers FCE information to be of complementary value versus those that do not. The information that a claimant is currently working might make the information from an FCE assessment appear less valuable, and thus influence the IP's perception of the complementary value of FCE information. Functional disability was also assessed with the revised Oswestry questionnaire. The revised Oswestry questionnaire is derived from the Oswestry questionnaire [18] and is a 10-item instrument designed to measure the effects of pain on functional disability. Results of the revised Oswestry questionnaire were noted in numbers of claimants according to the 5 classes outlined by the revised Oswestry questionnaire: 0-20%, 20%-40%, 40%-60%, 60%-80%, 80-100% (a higher class indicates a higher level of functional disability).

Differences were studied using independent t-tests for the relationship between work experience of IP and the outcome on the question about the complementary value of FCE information. Chi square tests were used to assess differences between the two groups - IPs who do or do not consider the FCE information to be of complementary value - on familiarity with FCE (IPs), location of disorder of the claimant, and claimant's work status. Kendall's tau-c was used to test for differences between the two groups of IPs regarding the scores of the revised Oswestry outcome of the claimants.

For the answers to the question about the change in IP judgment based on FCE information, the numbers and percentages of IPs in the three categories (IP's assessment remained unchanged, increased, or decreased with respect to the claimant's abilities) were noted for each of the 12 activities. In addition, these data and their correlation to whether the IPs did or did not consider the FCE information to be of complementary value was tested

using Chi square tests. The outcome of whether FCE information had reinforced the judgment of physical work ability was scored affirmatively when 66% of the IPs answered 'yes.'

Answers to the third question were noted as the number and percentage of IPs answering 'yes' or 'no' with regard to their intention to use FCE in future assessments, along with the reasons given for this intention and the groups of claimants for which FCE information was considered to be particularly useful. Furthermore, differences between the group of IPs who did or did not consider the FCE information to be of complementary value were tested with reference to the intention of future use of FCE information by using Chi square tests.

Finally, the relationship between the answers concerning complementary value and reinforcement of judgment and intention of future use were studied using independent t-tests. The significance level of all statistical tests was set at $p < .05$.

7.3 Results

Fifty-four IPs were prepared to take part in the study and signed an informed consent form, resulting in a response rate of 54%. For 26 of these IPs, no claimant application forms were received within the study period and they were not included in the study. This left 28 IPs, each with one claimant with MSD whose physical work ability was assessed. Table 1 shows descriptive information of the study population. The mean age and standard deviation (SD) of the IPs was 48 (7) years, and 64 % of the IPs were male. Their mean experience (SD) in the assessment of disability benefit claimants was 15 years (7). Fifteen of the 28 IPs were familiar with FCE. Between the two groups of IPs, those whose claimants did or did not enter the study, no significant differences existed for age, gender, or years of work experience. The claimants of IPs who were familiar with FCE preceding the study participated more often than claimants from IPs who were not familiar with FCE ($p = .02$).

Twenty of the claimants included were seen in the context of a disability re-assessment procedure, i.e. they were currently receiving a full or partial disability pension and were re-assessed pursuant to statutory requirements. The other eight claimants came for initial assessment of a disability claim after 24 months of sick leave. The 28 claimants were subjected to a standard Ergo-Kit test protocol by 13 certified raters at 13 locations throughout the Netherlands.

The mean age (SD) of the claimants was 46 years (5) and 41% of the claimants were male. Fifteen of the 28 claimants had MSDs of the neck and back, and eight had a disorder extending to more than one region. Upper and lower extremity disorders were reported in two and three claimants, respectively.

Table 1: Gender (number; percentage), age in years (mean; SD), years of experience (mean, SD) and familiarity with FCE (number; percentage) of the insurance physicians (N = 28). Gender (number; percentage), age in years (mean; sd), and region of disorder (number, percentage) of the FCE claimants (N = 28).

	Insurance physicians N = 28	Claimants N = 28
Men (number, percent)	18 (64)	11 (39)
Women (number, percent)	10 (36)	17 (61)
Age in years (mean, sd)	48 (7.4)	46 (4.7)
Experience in years (mean, sd)	15 (6.9)	
Familiarity with FCE (number, percent)	15 (54)	
Region of disorder (number, percent):		
Upper extremity		3 (7)
Lower extremity		2 (7)
Neck and Back		15 (55)
Combination		8 (30)

Complementary value

Nineteen of the 28 IPs (68%) indicated that FCE had complementary value for assessment of the physical work ability of the claimant under review. This is a greater proportion than the stated threshold of 66%. Only eight IPs gave a voluntary comment in addition to the response about complementary value. The tendency in the spontaneously given comments was that the complementary value of the FCE information was limited. Referring to our sub-question, neither work experience nor familiarity with FCE was significantly different between the group of IPs that did consider FCE information to be of complementary value and to the group of IPs that did not consider the FCE information of complementary value.

Change and reinforcement of judgment

The IPs indicated that they changed their judgment 127 of the 336 times (38%). In 209 (62%) cases, the IPs indicated no change in their judgment about the work ability of the claimants to perform the 12 activities because of the FCE information. In the two subgroups of IPs, the number of changed judgments about the ability to perform the 12 activities was 108 (47%) in the group of 19 IPs that considered the FCE information to be of complementary value and 19

(16%) in the group of nine IPs that did not consider FCE information of complementary value. In the latter group, more than 80% stuck to their judgment *versus* 53% in the group of IPs that considered the FCE information to be of complementary value. The difference between the two categories of IPs on this measure was significant (*p* value= .004). The numbers and percentages of IPs who changed their judgment after studying the FCE information, and the direction in which the judgment was changed for the 12 activities in question, are presented in Table 2.

Table 2: Numbers and percentages[*] of insurance physicians who changed their assessment of a claimant's ability to perform 12 different activities after studying FCE information, and the direction of this change.

	Change		More ability		Less ability	
	N	(%)	N	(%)	N	(%)
Walking	9	(35)	6	(67)	3	(33)
Sitting	9	(32)	5	(56)	4	(44)
Standing	9	(33)	5	(56)	4	(44)
Lifting/ carrying	15	(58)	7	(47)	8	(53)
Dynamic trunk movement	5	(20)	3	(60)	2	(40)
Static bending trunk	5	(19)	1	(20)	4	(80)
Reaching	7	(27)	1	(14)	6	(86)
Moving above shoulder height	10	(44)	2	(20)	8	(80)
Kneeling/ crouching	9	(36)	1	(11)	8	(89)
Repetitive movements hands	6	(28)	3	(50)	3	(50)
Specific movements hands	5	(23)	3	(60)	2	(40)
Pinch/ grip strength	6	(27)	3	(50)	3	(50)

Four IPs did not change their assessment for any activity. On average, IPs changed their assessment of four activities (mean 4.0, SD 2.6*)*, with a range of from 0 to 10 activities. About 58% of the IPs who indicated that they changed their judgment of the claimant's ability to lift and carry with seven raising their estimate and eight lowering it. Similarly, 44% of IPs changed their assessment of the ability to work above shoulder height, lowering their estimate in eight out of 10 cases. Eight IPs lowered their estimate of the ability to kneel or crouch after studying the FCE information, while only one raised it. Finally, 75% of the IPs who indicated that they altered their judgment about the ability to walk raised their assessment.

[*] Since not all IPs assessed all types of activity, the percentages are not all out of 28.

A majority of 76 % of the IPs (16 of the 21) indicated that FCE information reinforced their judgments of physical work ability. This is more than the stated threshold of 66%. Thus, we conclude that FCE information did serve to reinforce IPs' judgment in this study. Eleven of the 14 IPs (79%) from the group that considered FCE to be of complementary value and five of the seven (71%) that considered the FCE information not to be of complementary value, indicated that the FCE information had reinforced their judgment. The difference between the two groups was not significant.

Future use

Eighteen of the 28 IPs (64%) indicated that they intended to use information from FCE assessments in future disability claim procedures. Of this group, twenty IPs were positive and eight were negative about the complementary value of FCE information. Seventeen of the 20 IPs (85%) who were positive about the complementary value of FCE information indicated that they intended to make use of this information in the future. Only one of the eight IPs (13%) who was not positive about the complementary value of FCE information indicated that they intended to make use of this type of information in the future. Arguments given in favour of FCE information were: the information is objective, it gives a better insight in the claimant's work ability, and it leads to better acceptance of the IP's decision by the claimant. Arguments given against future use of FCE information were: the complexity of the FCE report, the duration and cost of an FCE assessment, the fact that FCE information does not make a distinction between restrictions in work ability based either on disorders or on personal traits, and that malingering was thought to be possible. The groups of claimants for which FCE information was thought to be useful were claimants with MSDs, claimants with medically unexplained disorders, claimants with complex disorders (which make it difficult to assess the work ability, like fibromyalgia, chronic fatigue syndrome, whiplash, and repetitive strain injury), and claimants with a large discrepancy between objective findings and subjective feelings of disability. Some IPs gave arguments in favour of FCE assessment not specifically related to claimant characteristics, like when the question about fitness for one's own job is at stake.

Complementary value & future use

The association between believing that FCE information has complementary value and those with the reported intention of using FCE information in future disability claim assessments was significant (p-value = .01), confirming the hypothesis that a positive judgment about the complementary value of FCE was related to an intention of future use of this information in disability claim procedures. No significant association was found between the answer about the complementary value and believing that their judgment was reinforced. This implicates that FCE information can reinforce the judgment about the physical work ability without being judged as of complementary value according to IPs.

7.4 Discussion

The aim of this study was to establish whether FCE information had complementary value for IPs in their judgment of physical work ability. More than two-thirds of the IPs affirmed the complementary value of FCE in this context, and stated that it helped to provide a firmer basis for their decisions. Sixty-four percent of the IPs indicated that they intend to include FCE information in future disability claim assessments.

In contrast to earlier studies about FCE information in work situations [19-23], this study took disability claim assessments into context. The strength of the study is that FCE information was introduced into the normal routine of disability claim assessments. This means that the IPs' judgment about the complementary value of FCE information was placed in the context of work ability assessment practice; it should be noted, however, that the FCE information did not influence the official judgment in the disability process.

When an instrument is stated to have complementary value for IPs in the assessment of physical work ability, it should reinforce their judgment and/ or alter their judgment of the physical work ability. A majority of IPs did, indeed, indicate that the FCE information had reinforced their initial judgment. Also, a majority of IPs altered their initial assessment as only four IPs stuck by their original appraisal of all activities considered. Three comments may be made in this regard:

i) Reinforcement of one's judgment does not necessarily exclude all changes in the assessment of individual aspects - An IP may well change his opinion about the claimant's ability to perform one or two activities while still feeling more confident in his initial appraisal of the overall physical work ability.

ii) IPs did not change their opinion in any specific direction in this study. Roughly equal numbers revised their estimates upwards versus downwards. This is in contrast to the results of a previous study that compared impairments in work ability as reported by the claimant, as assessed by the IP, and as estimated by FCE assessments. In this study, it was found that the self-reported level of impairment was highest, that derived from the judgment of IPs was at an intermediate level and that derived from FCE assessment was in general lowest [23], indicating that FCE would generally result in a downward revision of assessed impairment. The present study did not show such a shift towards higher work ability assessments (lower impairment assessments) after the IP had studied the FCE results.

iii) No systematic connection was found between the location of the disorder (upper or lower extremity) and the reported changes in the assessment of performance. For instance, the ability to reach and perform activities above shoulder height, may be seen as a potential impairment in workers with upper extremity disorders, but was altered as well in claimants with disorders of the back or lower extremity.

To determine what factors might give cause to the opinion of some IPs that FCE information is of complementary value for the judgment of physical work ability in disability claim assessments, we examined differences between the groups of IPs that did or did not consider the FCE information of complementary value. We analysed characteristics of both the IPs and of the included claimants. Work experience and familiarity with FCE were thought to be aspects that have influence on the outcome of complementary value of FCE. However, this did not appear to be the case. The other IP characteristics were not different, either. Although there was a difference in familiarity with FCE and participation of claimants in the study, there was no relationship between this finding and the outcome with regard to the question about complementary value, and therefore, the difference is not relevant to the question posed by this study. Another possible explanation for the difference between the two groups of IPs could result from a difference in their claimant population. Again, the different characteristics that were examined, location of disorder and work status, showed no significant differences between the two subcategories of IPs. The results of the Revised Oswestry Questionnaire were also not found to correlate with the judgement of the IPs about the complementary value of FCE. Therefore, it remains unclear why IPs have different opinions about the complementary value of FCE information. Regardless, however, whether or not FCE information is of complementary value influences the intention of future use. Thus, the hypothesis that when IPs consider FCE information to be of complementary value, they would also intend to make use of this information in future disability claim assessments,

appears to be correct. One explanation for this might be that IPs do not have many instruments upon which to base their judgment when assessing work ability of claimants in the context of disability claims. FCE information is a potential instrument to assist them in this task. IPs in the group that considered the FCE information to be of complementary value, changed their judgment significantly more often compared with their colleagues with the opposing opinion. The group who believed FCE information was not of complementary value remained at their previous judgment more often.

The following remarks may be made with regard to the external validity of the results: *i*) In this study, IPs could not directly refer claimants for FCE assessment; moreover, claimants were completely free to decide whether they would participate and undergo the FCE test. This avoids the possibility of bias present in cases where claimants are referred to assessments like FCE by IPs. Since the IPs could not refer the claimants for FCE, their positive appraisal of the complementary value of such tests is unlikely to be falsified by their preconceived views. *ii)* Since a majority of the IPs indicated that they would consider using FCE information in future disability claim assessments, it may be expected that if they could refer claimants for FCE assessment in appropriate cases, their appreciation of the complementary value of FCE information might be even higher.

IPs believe that claimants for whom a discrepancy is found between the subjective complaints and expected objective findings would be a suitable target group for FCE in future disability claim assessments . In these cases, the claimant, who is usually the primary source of information [2], will naturally tend to be a low estimate of their own physical work ability. The findings from physical examination, on the other hand, usually show little or no objective abnormality findings and cannot support the patients' view of their work ability. Whether this patient group is, indeed, a more suitable group for these forms of assessment of physical disability cannot be concluded from this study. This would, however, be an interesting topic for future research.

Some remarks are necessary about the choice of tests. In our study, we used the full FCE Ergo-Kit. Since the objective was to investigate the complementary value of FCE information for IPs in assessment of the work ability of claimants with MSD, there is no reason to limit the extent of the test battery. It is conceivable, however, that not all information generated by a full FCE may be required in all situations. It may not be relevant, for example, to assess the ability to kneel and crouch in claimants with impairments of the upper extremities. There have been requests for shorter FCEs, more specifically aimed at the work that the disabled worker is expected to do [24] or targeting the specific impairment in regional disorders [25,26]. However,

this study shows clearly that FCE information leads IPs to change their judgment even on activities not directly related to the underlying disorder and that IPs still regard this information as having complementary value. This is an argument for continuing the use of full FCEs. It is also noteworthy that the groups of claimants in whose assessment IPs indicated that FCE information would form a useful supplement largely presented problems of general physical functioning. Use of a full FCE would therefore seem to be called for in the assessment of such cases.

Finally, the practical implications of this study should be discussed. The positive evaluation of FCE information expressed by IPs in the study population argues for the introduction of FCE as a part of the disability claim assessment procedure, especially for those groups of claimants for which IPs think that FCE information yields maximum results. However, this study is based solely on the judgment of IPs towards the complementary value of FCE information. The prognostic value of FCE as a routine instrument in disability claim assessments has yet to be established.

ACKNOWLEDGEMENTS: We would like to thank all functional capacity raters, insurance physicians and claimants who participated in this study.

Reference List

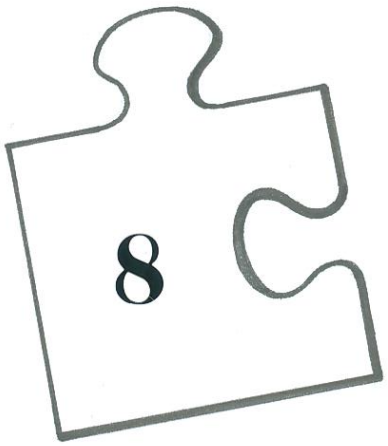
1. Van de Mheen H, Stronks K, Schrijvers CTM, Mackenbach JP (1999) The influence of adult ill health on occupational class mobility and mobility out of and into employment in The Netherlands. Soc Sci Med 49: 509-518
2. Bont de A, Brink van den JC, Berendsen L, Boonk M (2002) Limited control of information for work disability evaluation [De beperkte controle van de informatie voor de arbeidsongeschiktheidsbeoordeling: in Dutch] Ned Tijdschr Geneeskd 146: 27-30
3. Rainville J, Pransky G, Indahl A, Mayer EK (2005) The physician as disability advisor for patients with musculoskeletal complaints. Spine 30: 2579-2584
4. Wind H, Gouttebarge V, Kuijer PPFM, Frings-Dresen MHW (2005) Assessment of functional capacity of the musculoskeletal system in the context of work, daily living, and sport: a systematic review. J Occup Rehab 15: 253-272
5. Statistics Netherlands (2004) http//www.cbs.nl/theme labour, income and social security [in Dutch]
6. Picavet S, Schouten JSAG (2003) Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC_3- study. Pain 102: 167-178
7. Le Pen C, Reygrobellet C, Gérentes I (2005) Financial cost of osteoarthritis in France. The COART France study. Joint Bone Spine 72: 567-570
8. Lubeck DP (2003) The costs of musculoskeletal disease: health needs assessment and health economics. Best Pract Res Clin Rheum 17: 529-539
9. Lyth JR (2001) Disability management and functional capacity evaluation: a dynamic resource. Work 16: 13-22
10. Harten JA (1998) Functional Capacity Evaluation. Occup Med 13: 209-212
11. Vasudevan SV (1996) Role of functional capacity assessment in disability evaluation. J Back Musculoskel Rehab 6: 237-248
12. Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2006) The utility of functional capacity evaluation: the opinion of physicians and other experts in the field of return to work and disability claims. Int Arch Occup Environ Health 79: 528-534
13. Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2005) Intra- and interrater reliability of the Ergo-Kit Functional Capacity Evaluation method in adults without musculoskeletal complaints. Arch Phys Med Rehabil 86: 2354-2360

14. Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2006) Reliability and agreement of 5 Ergo-Kit Functional Capacity Evaluation in lifting tests in subjects with low back pain. Arch Phys Med 87: 1365-1370
15. Hudson-Cook N, Tomes-Nicholson K, Breen A (1989) A revised Oswestry disability questionnaire. In: Roland M, Jenner JR, eds. Back Pain: New approaches to Rehabilitation and Education. Manchester: Manchester University Press, page: 187-204
16. Razenberg PPA (1992) Formation of judgment: scientific framework [Verzekeringsgeneeskundige oordeelsvorming:inzicht in de praktijk: in Dutch] PhD thesis. University of Amsterdam, Amsterdam.
17. Kerstholt JH, de Boer WEL, Jansen EJM, Bollen D, Rasker PC, Cremer R (2002) Psychological aspects of disability claim assessment. [in Dutch] TNO report: TM-02-C051. Hoofddorp. Page 33
18. Fairbank JCT, Couper J, Davies JB, O'Brien JP (1980) The Oswestry low back pain questionnaire. Physiotherapy; 66: 271-273
19. Gross DP, Battié MC, Cassidy JD (2004) The prognostic value of Functional Capacity Evaluation in patients with chronic low back pain: part 1; timely return to work. Spine 29: 914-919
20. Gross DP, Battié MC (2004) The prognostic value of Functional Capacity Evaluation in patients with chronic low back pain: part 2; sustained recovery. Spine 29: 920-924
21. DP, Battié MC (2005) Functional Capacity Evaluation performance does not predict sustained return to work in claimants with chronic back pain. J Occup Rehab 15: 285-294
22. Gross DP, Battié MC (2006) Does Functional Capacity Evaluation predict recovery in workers' compensation claimants with upper extremity disorders? Occup Environ Med 3: 404-410
23. Brouwer S, Dijkstra PU, Stewart RE, Goeken LN, Groothoff JW, Geertzen JH (2005) comparing self-report, clinical examination and functional testing in the assessment of work-related limitations in patients with chronic low back pain. Disabil Rehabil 27: 999-1005
24. Frings-Dresen MHW, Sluiter JK (2003) Development of a Job-specific FCE protocol: the work demands of hospital nurses as an example. J Occup Rehabil 13: 233-248

25. Gross DP, Battié MC, Asante A (2006) Development and validation of a short-form Functional Capacity Evaluation for use in claimants with low back disorders. J Occup Rehab 16: 53-62
26. Soer R, Gerrits EHJ, Reneman MF (2006) Test-retest reliability of a WRULD functional capacity evaluation in healthy adults. Work 26: 273-280

Chapter 8

General Discussion



8.1 General Discussion

This thesis focused on the utility of information from FCE tests for insurance physicians' (IPs) performance of the statutory task of assessing the ability of disability benefit claimants to work. IPs do not have many instruments at their disposal to support them in this task. Since FCE methods were developed to supply information about the physical work ability, they may support IPs in this task. The first part of this chapter summarizes the main findings on the utility of FCE and contains methodological considerations of the studies as they are presented. The second part discusses the place of FCE methods in relation to assessments of physical work ability. Finally, the implications for assessments of work ability in the context of disability claims in relation to FCE are discussed. The chapter will conclude by answering the main research question, discussing the implications of the study results in the process of disability claim assessments and providing recommendations for future research and practice.

8.2 Summarizing the main findings

The main research question addressed by this thesis, which concerns the utility of FCE in the assessment of physical work ability by IPs, was addressed by studying six sub-questions. The first of these asked what instruments are available that can be used to assess the physical capacity of people with MSD to perform the various activities required in the context of work, sport and daily life, and what can be said about the reliability and validity of some of these instruments. The answers to this question, as revealed by a systematic review of the literature were given in Chapter 2, and may be briefly summarized as follows. A number of questionnaires and functional tests have been developed to assess physical functional capacity in specific contexts. Four questionnaires met the reliability and validity criteria set, viz: The Oswestry Disability Index [1], the Pain Disability Index [2], the Roland-Morris Disability Questionnaire [3] and the Upper Extremity Functional Scale [4]. No functional test was found that could meet the criterion for reliability or validity. The second sub-question was: What is known about the reliability and validity of FCE methods available in the Netherlands. The results of a systematic review of the literature were presented in Chapter 3. Twelve papers were identified for inclusion and assessed for their methodological quality. The conclusion was that more rigorous studies are needed to demonstrate the reliability and validity of FCE methods. The third sub-question asked what the reliability and agreement is of 5 EK FCE lifting tests in subjects with low back pain. Chapter 4 describes the results of this study. Five EK FCE lifting tests (two isometric and three dynamic lifting tests) were studied. There

appeared to be good reliability and agreement between raters of the isometric and dynamic EK FCE lifting tests in subjects with low back pain. The fourth sub question was: how do return-to-work case managers and disability claim experts perceive the utility of FCE for their work and what arguments do they present with respect to the utility (Chapter 5)? The main conclusion was that return-to-work case managers tended to regard FCE as more useful than disability claim experts. Arguments given in favour of the utility of FCE were its ability to confirm one's own opinions and its objectivity. Arguments against it were the redundancy of information provided by FCE and the lack of objectivity. Another rather surprising finding was that very few of the experts that were questioned appeared to be familiar with FCE from their own experience. The last two sub-questions were involved analysis of data from a study described in Chapters 6 and 7. The fifth sub-question, which is in Chapter 6, was: Does information derived from FCE tests lead an IP to change his assessment of the physical work ability of a disability benefit claimants with MSD? It was found that IPs changed their assessment of the ability of claimants with MSD to perform work-related activities significantly more often after they received FCE information than when they received no FCE information. The changes in judgment were mostly in line with the FCE results, in cases of more or less physical work ability. The sixth sub-question was: Is information derived from FCE tests of complementary value to IPs in their assessment of the physical work ability of disability benefit claimants with MSD? This question was studied in Chapter 7. It was found that a majority of IPs considered FCE information to be of complementary value in this context. The threshold of 66% of affirmative answers which was required for significance was exceeded. A significant majority of IPs (76%) also indicated that the FCE information reinforced their judgment. IPs who considered the FCE information to be of complementary value, indicated significantly more often that they had changed their judgment about the listed activities compared to the IPs who did not consider this information to be of complementary value. An ample majority of 76% of the IPs intended to use this information to deal with future disability claims. It was found that there was a significant association between the believe that the FCE information was of complementary value and the expressed intention to use this type of information in future disability claims. On the basis of the results presented in Chapters 6 and 7, it may be concluded that FCE information is indeed useful to IPs who have to assess the physical work ability of persons with MSD in the context of disability benefit claim procedures.

8.3 Methodological considerations

The main research question posed in this thesis concerns the utility of FCE to IPs for the assessment of the physical work ability of claimants with MSD in the context of statutory long-term disability claim assessments. First of all, there is the question why was chosen for the EK FCE as the instrument of which the complementary value for the assessment of physical work ability by IPs was studied. From the review study in Chapter 2 can be concluded that there are a number of questionnaires that are both reliable and valid for assessing physical work ability. The most important reason for not choosing these questionnaires is that the information in these questionnaires is not performance based. This information received from the patient regarding his ability to perform activities is comparable to the information that an IP receives in the anamnesis from the patient. In addition, the Upper Extremity Functional Scale [4] is a specific questionnaire to assess the physical work ability of the upper extremity and is therefore not fit for a judgment of the total physical work ability. The three other questionnaires are aimed at assessing the physical work ability more in general, though with a focus on low back complaints.

This study is about the influence of a different way of assessing the physical work ability, i.e. by using a performance based test. A study designed to investigate the question what the utility of FCE information for IPs at assessing the physical work ability of claimants with MSD in the context of statutory long-term disability claim assessments must cover all its aspects: the overall context of statutory disability claim assessments and the procedure for which the utility of FCE is to be studied. Furthermore, it is stipulated that two groups of participants are involved: IPs and patients with MSD. However, before the discussion can focus on these aspects, it is necessary to pay attention to the clinimetric qualities of the used instrument in this thesis: the EK FCE. As stated in the introduction, safety, reproducibility (reliability and agreement), validity, utility, and practicality are the main aspects to consider. The safety of the EK FCE can be considered to be sufficient, as was argued in the Introduction. Before an instrument like the EK FCE can be used legitimately, its resultant information should be demonstrated to be reliable. Chapter 3 concluded that there were no rigorous studies on the reliability of EK FCE. Lifting tests of the EK, both static and dynamic, were subject of reliability studies. Importantly, these tests were found to be reliable. Although the reliability of the total EK FCE is not proven, these studies give no suggestion that the reliability of the EK FCE is inadequate. Reliability was also found to be equally good in another FCE method: the IWS [5,6]. EK FCE was selected for this study because of its

availability throughout the Netherlands, an important argument because it enables performance of the study nationwide in the normal procedure of disability claim assessments. Nevertheless, when the EK FCE is used in disability claim assessments, the reliability of the other test attributes should also be studied. In the introduction, it was explained that the validity of FCE and, more specifically, of the EK FCE is an important issue. There is some evidence that the EK FCE is valid. There is sufficient proof of the face validity of the EK FCE, since the test procedures are fully described in a manual and they are standardized. In addition, the procedure for drawing up a report is specified and the test leaders are certified. The activities of the test are derived from activities mentioned in the Dictionary of Occupational Titles (DOT)[7]. This means that there is a direct link between the activities and work demands and, therefore, these activities can be considered to be work- related. According to Portney and Watkins [8], content validity refers to the adequacy with which an instrument can cover all the parts of an underlying universe of content and reflects the relative importance of each part. To have sufficient content validity the EK FCE should measure all the important aspects of physical work ability, but be free from the influence of factors that are irrelevant to the purpose of the measurement. Through the relation with the DOT physical demands, there is some content validity for a number of FCE assessment methods [9]. There are no specific studies on the content validity of the EK FCE, but the parallel between the EK FCE and other FCE methods in relation to the DOT leads to the conclusion that there is proof of some content validity for the EK FCE. Other forms of validity of the EK FCE need to be studied, but this is beyond the scope of this thesis, which focuses on the utility of FCE information for IPs. Utility is an aspect that refers to the user of the EK FCE information. To that end, it was determined whether the information had a complementary value for its intended purpose. In case of IPs, the purpose of the information is the assessment of physical work ability in the context of disability claim assessments. The influence of the information from the EK FCE instrument on the judgment of the physical work abilty by IPs is central to this thesis.

As explained in Chapter 1, the assessment of physical work ability resembles the medical diagnostic process of disease. It follows that studying the utility of FCE for the assessment of physical work ability is similar to studying the utility of a diagnostic instrument. The usual procedure in such studies is to define a ‘golden standard’ for the diagnostic process and to compare the new instrument with this golden standard. Sensitivity and specificity are the appropriate outcome measures in that situation. There is not a golden standard for the assessment of work ability; however, ‘the proof of the pudding is the eating of the pudding’,

meaning that the true work ability is present when people can work day by day without deterioration of their physical ability and within the limits of the physical work ability set by the IP. The IP assesses the work ability in general – largely on the basis of information received from the claimant - and physical work ability is a part of this general work ability. The IP's judgment of the physical work ability is the outcome of the assessment process and therefore, should be tested to determine the diagnostic value of FCE information. Two aspects need to be studied in this respect: the extent to which the FCE information causes the IP to change his assessment of the physical work ability and the extent to which it reinforces his judgment. To address the first issue, IPs performed two assessments of the ability of claimants to perform twelve work-related activities and recorded the results on visual analogue scales (VAS). The claimants were divided into two groups, an experimental and a control group. In the experimental group, FCE information about each claimant was provided before the second assessment, while no additional information was provided in the control group. The shift in the IPs' judgment between the two assessments was calculated for the two groups. To address the second issue, the IPs were asked whether they regarded FCE information as having complementary value for their assessment of physical work ability. This study of the effect of FCE information on IPs' judgment was carried out under normal working conditions and not in an artificial setting. The outcome of the study is thus directly related to the actual process of disability claim assessment in the Netherlands. The execution of the study was subject to certain constraints. The time between the first assessment of physical work ability by the IP and the FCE tests varied, but it was more than six weeks on average. This is a long period, and there is a certain risk that the claimants' physical medical condition may have changed during this interval. It should be noted, however, that the claimants who participated in the study had all been on sick leave for at least two years, and often much longer. They were all suffering from chronic disorders such that a dramatic change in their physical medical condition was not to be expected in this time period. Since the IP's review of the claimant's work ability was only based on inspection of the claimant's file, the time between the first and the second assessment was less relevant. In fact, this long delay has the advantage of eliminating the risk of recall bias. There was not a significant difference in the amount of time that elapsed from the first assessment and the second by the IP between the two groups.

Another important methodological consideration is the choice of the VAS system for recording of the IPs' assessment of the physical work ability of claimants. Currently, the instrument that is routinely by IPs for recording physical work ability in the context of disability claim processing in the Netherlands is the Functional Ability List (FAL). The main

purpose of the FAL is to facilitate the selection of suitable jobs for claimants found to have residual work ability. The FAL rates physical work ability on an ordinal scale in 2, 3, or 4 categories, and will therefore not reflect relatively small changes as well as VAS. This is a disadvantage for the purposes of the present study, where the shift in IPs' judgment is one of the main points under investigation. We did not determine whether the judgment of the IP on the VAS scale was a true account of the actual physical work ability for that activity, rater, we determined whether there was a shift in judgment under the influence of FCE information.

An instrument was needed that was sensitive enough to register this shift, such as the VAS scales[10,11]. In addition, a pilot study was performed to ensure that the VAS scales were a feasible instrument in the context of disability claim assessments by IPs. This led to two alterations in the original scheme, viz. adjustment of the definitions of end points of the VAS scales, and determining that a 1 cm shift on the VAS scale indicated an intended change in judgment by the IP of the claimants' ability to perform that activity. Furthermore, VAS scales have been proven to be reliable and valid [10,11]. This is in contrast to the FAL as an instrument for recording physical work ability, of which no studies about reliability and validity for recording the physical work ability are found. There are no studies on the reliability of the FAL for recording physical work ability. Using VAS scales instead of the FAL also means that IPs could give an unbiased judgment of the physical work ability of the claimants involved, in the sense that this judgment is not directly related to the judgment of work ability in the statutory claim situation. The choice of VAS rather than the FAL as the method for recording IPs' judgment probably yields, therefore**,** a better estimate of the IP's real perception of the physical work ability of the claimants for the selected activities.

8.4 FCE and the assessment of physical work ability

In Chapter 1, work ability assessment was described as a diagnostic process, where uncertainty about the outcome plays a central role. Information is used to reduce uncertainty, both in problem solving and in medical decision-making. When applied to the assessment of physical work ability in the context of the processing of disability claims, this means that information should reduce the uncertainty concerning the true physical work ability. The present study cannot provide a direct answer to the question of whether FCE information actually lowers the level of uncertainty concering the true physical work ability, since the prior chance of ascertaining the true physical work ability is unknown. What can be concluded, however, is that the IPs in majority valued the FCE information to be of complementary value, that the information strengthened their judgment, and that the IPs that

considered the FCE information to be of complementary value altered their judgment in majority in the direction of the FCE outcome. All these findings tend towards a possible lowering of the level of uncertainty, but the quantity of the lowering of uncertainty cannot be measured. The assessment of physical work ability is a complex task because many factors can influence this variable. The ICF (International Classification of Functioning, Disability and Health), which was drawn up by the World Health Organisation offers a useful framework in which all these factors and their interdependence can be displayed. [12]

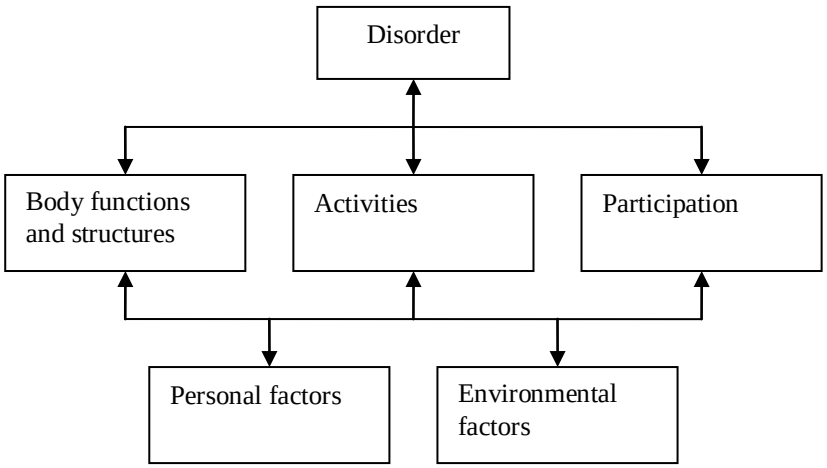


Figure 1: The ICF model (ref: ICF Geneva, 2001)

In another expert poll was studied what aspects of the ICF IPs focus on when handling disability claims of persons with MSD. The aspects found were body functions and structures, and participation [13]. Body functions and structures are considered in the anamnesis that occupies a central place in the IP's judgment about the physical work ability of a claimant with MSD. Participation is an important factor in disability claim handling because restoration of the ability to participate in work may be said to be the *raison d'être* of the disability claim assessment procedure. FCE offers a means of testing the claimant's ability to perform work-related activities. The information derived from FCE tests, being performance-based, is different in kind from the other types of information available to the IP.

When is an instrument useful in a diagnostic process that is used for work ability assessment? The different dimensions of utility are discussed in Chapter 5, where a distinction was drawn between utility for the organization, utility for the individual and intrinsic utility. The information provided by a diagnostic instrument is useful to an individual when it fills in gaps in his knowledge or reinforces his understanding of what was already known [14]. When this line of reasoning is applied to IPs who are assessing physical work ability in the context of

disability claims, this raises the question of what sources of information are commonly used by IPs during their appraisal of the physical work ability of claimants. It has been found that the most important source of information for Dutch IPs in disability claim assessments is the claimant himself [15]. This might explain why so much emphasis is placed on the recording of claimant information in the Netherlands. The key information received from the patient concerns the day-to-day activities he/she performs and the limitations he/she experiences in performing them [16]. This information is important because it reflects not only the claimant's functional performance but also his ability to participate in daily life and work. These two aspects, of functional performance and participation, represent the focus of the IP's attention in disability claim assessments [13]. It should not be forgotten, however, that the context of the assessement is a procedure in which the claimant is seeking financial compensation for his loss of functional physical ability to perform what used to be his normal work. Thus, there may be information bias on the part of the claimant. Several studies have shown that workers cannot accurately judge the exposure to physical activities in their job at a detailed level [17-19]. Hence, basing the assessment of physical work ability solely on information received from the patient might lead to an incorrect outcome. Other types of information commonly available to the IP, such as statements from the physicians treating the claimant, X-ray photograhs, and the results of blood tests etc., focus on diagnosis, severity of the disorder, treatment and prognosis – i.e., on body functions and structures in terms of the ICF model. They do not directly address the claimant's functional performance and ability to participate in day-to-day life and work. FCE tests show how claimants perform over a limited period of time in a simulated work situation which contributes to the evaluation of the ability to perform work-related activities. As mentioned above, this is an important aspect of IPs' judgment of physical work ability in disability claims. In this respect, the performance-based FCE information differs in this respect from the information about the physical work ability from the other sources. The view expressed by IPs in the study on the utility of FCE described in Chapter 5 is that information derived from FCE tests is objective as it is performance-based in nature[14]. Several authors have referred to this aspect of FCE assessment methods [20-22].

In conclusion, FCE information is useful because IPs indicate that the information is of complementary value, FCE information strengthens the judgment of physical work-ability, and FCE information results in alterations of the IPs judgments about the claimant's ability to perform the physical activities.

8.5 Implications for disability claim procedures

Disability claim procedures are developed with the objective of implementing a government's social insurance policy. Therefore, the legislator has made them subject to rules and conditions that are not strictly medical in nature but also reflect the constraints of the insurance system and the government's underlying policy considerations. IPs are employed by the agency charged with implementation of this insurance system and have a statutory duty to ensure that their judgments comply with all these rules and conditions. One of the fundamental basic assumptions is that the same rules must apply to all claimants. But, IPs are also physicians and, like their colleagues in hospitals and general practices, they have a duty to promote the health and welfare of the patients they see in the course of their work. Thus, there are always two sides to any disability claim assessment**:** the legal side and the health care side. FCE tests can only play a role in disability claim assessment if the information they provide is useful to the IPs performing these assessments and is in line with the medico-legal setting. The present study shows that the first of these conditions is met, since IPs consider FCE information to be useful for their assessment of physical work ability. The medical aspects are also satisfied because FCE tests measure the ability to perform work-related activities safely. The IP also has a statutory duty to ensure that the legal aspects are met. He must determine whether the information provided by the FCE tests about physical work ability fits in with all his other considerations about the possibility of the claimant's returning to work (if necessary adapted to his limited capacities), which the legislator stipulates is desirable – even mandatory – in the absence of overriding contra-indications. The FCE information only has a complementary value to IPs because, while it may be valuable in helping the IP to come to a decision on the issue of physical work ability, it can never replace his judgment concerning the claimant's work ability in a wider context. By indicating that FCE provides information of complementary value, the IPs implicitly stated that there are no impediments to the inclusion of this information in disability claim procedures where physical work ability is at stake. The FCE information can be combined with information from other sources to yield a broad, well-argued decision concerning the claimant's overall work ability. Since such decisions have far-reaching consequences for the lives of many people – not only the claimants but also their immediate family, etc. – disability claim assessments must meet high standards. FCE tests can be one of the resources used to back up the IP's judgment of physical work ability in disability claim assessments, and the inclusion of FCE information can lead to a judgment that is better argued.

8.6 General conclusion

It may be concluded on the basis of the study described in this thesis that the EK FCE has a good level of reproducibility (reliability and agreement) with regard to five lifting tests in people with MSD complaints. EK FCE does provide IPs with information that is useful for the assessment of physical work ability in patient groups with MSD complaints because it reinforces their confidence in their judgments, and can lead to significant shifts in assessed physical work ability. In addition, a majority of IPs indicate that they have the intention of using EK FCE information in future disability claim assessments. It follows that FCE is an appropriate instrument to support IPs in their assessment of the physical work ability of long-term disability claimants.

8.7 Future research

Although the main question posed in this thesis – is FCE useful in the assessment of physical work ability in the given context? – can be answered affirmative, several new questions emerged during the study. For example, what can be the complementary value if FCE information is brought into a different context, like return to work procedures for sicklisted workers with MSDs. Also the question whether the nature of the disorder is of importance to the question of complementary value is interesting. After all, there was a remarkable disparity between the views of the IPs who took part in the expert poll on the utility of FCE (Chapter 5) and those of the IPs who participated in the 'complementary value' study (Chapter 7) regarding the (groups of) claimants for whom they thought FCE information could be particularly useful. IPs from both groups mentioned claimants with medically unexplained disorders – but they had diametrically opposed views on them. The IPs from the expert panel were of the opinion that FCE was not useful for assessment of the work ability of this group, while the IPs in the complementary value study named this group as particularly suited for this type of assessment. At the moment no definitive explanation can be given for this difference in views.

Some of the IPs who took part in the studies described in this thesis commented that the utility of FCE information often depended on the specific context of the disability claims procedure, like the opinion that FCE information is not useful when there is a legal, or injury claim procedure. These concerns need to be dealt with before FCE can become part of the standard arsenal used in disability claim assessments.

In summary, it would seem that research on the following six topics could provide valuable information for all parties concerned: medical professionals, policy makers and FCE providers.

- Studies of the criterion and construct validity of EK FCE. Now that there is some proof of utility of FCE from the point of view of the user, further study on these qualities of the EK FCE instrument are needed. In this context, evaluation of the difference in FCE assessment outcomes and the complementary value of FCE information to IPs when different groups of claimants are subjected to FCE testing, could be useful.
- What is the complementary value of FCE information in disability claim procedures where IPs can refer 'suitable' claimants themselves for FCE testing? Moreover, what groups of claimants are likely to be considered suitable subjects in this case? What specific information from the FCE tests determine whether the information is of complementary value or not?
- Can FCE information contribute to reduction of interrater variability? Decreased variability would satisfy the statutory requirement that differences in outcome between comparable cases should be prevented as far as possible in the handling of disability benefit claims.
- How do claimants who have been subjected to FCE tests and have been informed of the test results value this information? What consequences does this knowledge have for their own estimate of functioning in their work- and daily life? The answer to this question should be of interest both to policy-makers and to IPs, since both these parties share a responsibility for the claimants who undergo work ability assessment.
- Can batteries of FCE tests be designed that are less time-consuming and less costly, while still remaining effective? The FCE tests currently available are time-consuming, and assess many different activities. If FCE testing is to become a routine part of disability claim assessment, it would be beneficial to devise shorter, more specific tests that retain the necessary reproducibility, validity and utility.

8.8 Recommendations

A number of recommendations may be made on the basis of the results of this study that, in specific instances, apply to IPs and policy-makers.

Recommendations for IPs

- Regardless of the decision-making aids that are introduced, uncertainty about the true work ability assessed in the context of disability claim procedures will remain. IPs should be aware of this uncertainty, and the compensatory design of disability claim assessments to handle the uncertainty. The ICF (International Classification of Functioning, Disability and Health) offers the necessary framework for this. Filling in all the components will lead to a more comprehensive assessment of functioning. This means that IPs must be made acquainted with other sources of information that can be used in the assessment of work ability, such as questionnaires and functional tests (e.g. FCE). They should learn how to interpret questionnaires and functional tests and how to use this information in disability claim assessment.

Recommendations for policy-makers

- IP employers should arrange and promote training courses in which IPs can learn how to use and interpret questionnaires and functional tests designed to provide information relevant to the assessment of work ability in disability claim procedures. Such training will help IPs to improve their assessment of work ability, thus reducing interrater variability and leading to better acceptance of their decisions by claimants. This might affect the amount of time an IP needs to handle a disability claim.
- The medical aspects of disability claim processing should be separated from the legal aspects; at present, the legal context of a disability claim assessment might bias the information of the patient. It is the claim of the patient on a disability pension that is assessed in a disability claim assessment. The moment of the disability claim assessment is legally fixed by law and it bears no relation to the claimant's medical condition. The patient's medical state is a dynamic variable that is characterized by periods of recovery, rehabilitation, deterioration, etc. while disability claim assessment is a static process. Assessment of the work ability in disability claim procedures should be targeted at helping the claimant to return to work. The moment at which this process is initiated is not linked to the time at which the statutory disability claim procedure is required to take place.

This thesis started by stating that the assessment of work ability in the context of long-term disability claim procedures is a complex matter, and the IPs who perform these assessments do not have many instruments to facilitate this endeavour. Assessment of physical work ability is akin to solving a jigsaw puzzle. Each additional bit of information helps to complete the picture, but some vital pieces may be missing. FCE is an instrument that has complementary value for the assessment of physical work ability, viz. adds a piece to the jigsaw puzzle. It might bring us closer to revealing the true picture of work ability.

Reference list

1. Fairbank JCT, Couper J, Davies JB, O'Brien JP (1980) The Oswestry low back pain questionnaire. Physiotherapy 66: 271-273
2. Tait RC, Chibnall JT, Krause S (1990) The pain disability index: psychometric properties. Pain 40: 171-182
3. Roland M, Morris R (1983) A study of natural history of low back pain. Part 1: Development of a reliable and sensitive measure of disability in low-back pain. Spine 8: 141-144
4. Pransky G, Feuerstein M, Himmelstein J, Katz JN, Vickers LM (1997) Measuring functional outcomes in work-related upper extremity disorders – Development and validation of the upper extremity function scale. J Occup Environ Med 39: 1195-1120
5. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN (2003). Test-retest reliability of the Isernhagen Work Sytems functional capacity evaluation in patients with chronic low back pain. J Occup Rehabil 12: 207-218
6. Reneman MF, Dijkstra PU, Westmaas M, Goëken LNH. (2002) Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. J Occup Rehabil. 12: 269-276
7. United States Department of Labor, Dictionary of Occupational Titles, (1991) 4th ed., US Government Printing Office, Washington DC
8. Portney LG, Watkins MP (2000) Foundations of clinical research. Applications to practice. 2nd ed. Chapter 6: validity of measurements, page 82-84. Prentice Hall Health Upper Saddle River, New Jersey
9. Innes Ev, Straker L (1999) Validity of work-related assessments. Work 13: 125-152
10. Wagner DR, Tatsugawa K, Parker D, Young TA (2007) Reliability and utility of a visual analog scale for the assessment of acute mountain sickness. High Alt Med Biol 8(1): 27- 31
11. Gallagher EJ, Bijur PE, Latimer C, Silver W (2002) Reliability and validity of a visual analog scale for abdominal pain in the ED. Am J Emerg Med 20 (4): 287-290
12. WHO (2001) International classification of functioning, disability, and health Geneva 2001
13. Slebus FG, Sluiter JK, Kuijer PPFM, Willems JHBM, Frings-Dresen MHW (2007) Work-ability evaluation: a piece of cake or a hard nut to crack? Disabil Rehabil 29 (16): 1295-1300

14. Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW (2006) The utility of functional capacity evaluation: the opinion of physicians and other experts in the field of return to work and disability claims. Int Arch Occup Environ Health 79: 528-534
15. De Bont A, Van den Brink JC, Berendsen L, Boonk M (2002) Limited control of information for work disability evaluation [De beperkte controle van de informatie voor de arbeidsongeschiktheidsbeoordeling: in Dutch] Ned Tijdschr Geneeskd 146(1): 27-30
16. De Boer WEL, Wijers JHL, Spanjer J, Van der Beijl I, Zuidam W, Venema A (2006) Discussion models in insurance medicine [Gespreksmodellen in de verzekeringsgeneeskunde: in Dutch] Tijdschr Bedr Verz Geneesk 14(1): 17-23
17. Wiktorin C, Karlqvist L, Winkel J (1993) Validity of self-reported exposures to work postures and manual material handling. Scand J Work Environ Health 19: 208-214
18. Campbell L, Pannett B, Egger P, Cooper C, Coggon D (1997) Validity of a questionnaire for assessing occupational activities. Am J Ind Med 31: 422-426
19. Van der Beek AJ, Frings-Dresen MHW (1998) Assessment of mechanical exposure in ergonomic epidemiology. Occup Environ Med 55: 291-299
20. Lyth JR (2001) Disability management and functional capacity evaluations: a dynamic resource Work 16: 13-17
21. Vasudevan SV (1996) Role of functional capacity assessment in disability evaluation J Back Musc Rehabil 6: 237-248
22. Mooney V (2002) Functional capacity evaluation. Orthopedics 25 (10): 1094-1099

Summary

The assessment of work ability in the context of long-term disability claim procedures is a complex matter, and the insurance physicians (IPs) who perform these assessments do not have many instruments to help them in this endeavour. As explained in the General Introduction (Chapter 1), the assessment of physical work ability bears a resemblance to medical diagnosis, with the impediments to good functional performance as the condition to 'diagnose'. Uncertainty about the true nature of the underlying condition is inherent in any diagnostic process. IPs try to minimize the uncertainty concerning the work ability of disability benefit claimants by a process of hypothesis testing based on the information they have collected. Much of this information comes from the claimant himself, who is asked by the IP to describe his ability to perform certain work-related activities and the limitations on this ability. A more objective approach might be to get the claimant to perform certain work-related activities in a simulated work situation, and to measure the results. This is the essence of functional capacity evaluation (FCE): FCE makes use of a standardized set of tests designed to measure performance in work-related activities. The reliability and validity of FCE have been the subject of several studies, and such studies are still ongoing. Almost all those studies were in the context of rehabilitation and return to work, however. This thesis focuses on the utility of FCE in disability claim assessment. The following research question may be formulated in this connection:

- What is the utility of FCE for the assessment of the physical work ability of a claimant with a musculoskeletal disorder (MSD) by an IP in the context of statutory long-term disability assessment?

Six sub-questions derived from this main question are addressed in Chapters 2 to 7.
In the chapters two and three, two systematic reviews of literature are presented. In the first of these two systematic reviews (Chapter 2) the search was aimed at instruments (questionnaires and tests) that can be used to assess the physical capacity of the musculoskeletal system in the context of work, daily life and sport. Studies of such instruments were included in the review if the authors specified the context in which the instrument was used. Thirty-four studies met this criterion. Four questionnaires, the Oswestry Disability Index, the Pain Disability Index, the Roland-Morris Disability Questionnaire and the Upper Extremity Functional Scale, were found to have high levels of reliability and validity. None of the functional tests studied scored high on both reliability and validity. The conclusion of this chapter was that it was best to combine a questionnaire and functional test in order to obtain a more comprehensive

assessment of physical work capacity. The second of these systematic reviews of literature (Chapter 3) focused on studies about the reliability and validity of four FCE methods: Blankenship system (BS), Ergos work simulator (EWS), Ergo-Kit (EK) and Isernhagen work system (IWS). The research in five databases resulted in 77 potential relevant studies but only 12 papers were included and assessed for their methodological quality. Both the interrater reliability and the predictive validity of the IWS were found to be good. However, the procedure in the intra-rater reliability (test-retest) studies of the IWS was not rigorous enough to allow any conclusion. No study was found about the reliability of the EWS, EK or BS. The concurrent validity of the EWS and EK was not demonstrated and no validity study was found about the BS. Conclusion of this chapter was that more rigorous studies were needed to demonstrate the reliability and validity of FCE methods, especially the BS, EWS and EK. Conclusion of these two reviews of the literature was that some questionnaires were reliable and valid enough to use in procedures about assessing the physical work ability, but functional tests and in particular FCE methods were not rigorously enough studied to comply with standards of reliability and validity. More studies primarily on the reliability and later also on validity of FCE methods are needed. As a first step, the reliability of EK FCE lifting tests was studied.

Chapter 4 is devoted to the study about the reliability and agreement of 5 EK FCE lifting tests in subjects with low back pain. Twenty-four patients with low back pain were included from physiotherapy centers and assessed by two raters at two different times with a three days interval and in a counterbalanced order. The five lifting tests consisted of two isometric lifting tests and three dynamic lifting tests and were derived from the EK FCE. Reliability was expressed as an intraclass-correlation coëfficient and agreement with a standard error of measurement (SEM). The mean interrater reliability of the both isometric strength test was high (.97 and .96) as well as for the three dynamic lifting tests (.95 for both the carrying lifting strength test and upper lifting strength test and .94 for the lower lifting strength test). The SEM varied between 1.9 and 8.6 kg. This suggests a sufficient level of agreement of the EK lifting tests. In conclusion, the results suggested that the reproducibility (ie, reliability and agreement between raters) of 5 EK lifting tests in subjects with low back pain was good. After studying the reliability of the EK FCE, the study continues with another aspect of an instrument that is used for diagnostic purposes, viz. the utility of the information for the user. As a first step, the view of experts who were familiar with FCE information in their work setting was studied.

Chapter 5 describes the results of an expert poll on the perceived utility of FCE as an instrument supporting return to work and disability claim assessment. Twenty-one return to work (RTW) case managers and 29 IPs working as disability claim (DC) assessors were interviewed by telephone using a semi-structured interview protocol developed for the purposes of this study. They were asked how they perceived the utility of FCE for their work, the arguments they presented for considering FCE useful or otherwise and the conditions they set for the use of FCE in their work. To be included in this poll, the respondents had to have experience of the use of FCE information in their personal work. RTW case managers rated the utility of FCE at a mean of 6.5 (SD 1.5) on a scale of 0-10. The DC assessors rated the utility of FCE somewhat lower on average, and their responses showed a wider spread: mean 4.8, SD 2.2. Arguments presented in favour of the utility of FCE were its ability to confirm the respondent's own judgment and its objectivity. Arguments against were the redundancy of the information FCE provides and its lack of objectivity. The indications for FCE testing were MSD, a positive self-perception of the patient about their work-ability, and the presence of an actual job. The contra-indications mentioned for FCE testing were medically unexplained disorders, a negative patient self-perception of work-ability, and the existence of disputes and legal procedures. Conclusion of this expert poll was that RTW case managers had a more positive view on the utility of FCE for their work setting than DC assessors. FCE appeared to be relatively unknown as an instrument for assessment of physical work ability in the groups of experts in the study. This leads to the question how this type of information is valued in practice in disability claim assessments by DC assessors of patients with MSDs.

In chapter six and seven, results of two studies in the same study group are described. In Chapter 6 the results are presented of a pre/post-test experimental study within-subjects of the effect of FCE information on the judgment of IPs concerning the physical work ability of claimants with MSD**.** A total of 100 IPs were randomly selected from the pool of 566 IPs employed by the Institute for Employee Benefit Schemes (UWV), who perform disability claims in the Netherlands. Fifty-four of these IPs complied with the inclusion criteria and signed an informed consent. Two claimants with MSD seen by each IP could participate in the study. First of all, during the regular disability claim assessment the IP scored the ability of both claimants to perform twelve work-related activities compared with their ability before the onset of disability, using a 10-cm visual analogue scale (VAS) to record the results. One anchor point of the VAS scale corresponded to complete inability to perform the activity in question compared to the situation before the onset of disability, while the other corresponded

to the ability to perform the activity at the same level as before the onset of disability. The first claimant of each pair, underwent FCE testing while the other served as a control. The FCE report was added to the claimant's file. Finally, the IP reviewed the physical work ability of both claimants, based solely on the contents of their medical files, and filled in the relevant VAS scores. The number of shifts of more than 1 cm in the VAS scores for the twelve work-related activities between the first and the second assessment was counted, and served as a measure of the shift in the IP's judgment. The McNemar Chi square test for paired binomial data showed a significantly greater shift in the IP's assessment of the physical work ability of claimants with MSD in the experimental group than in the control group. The majority of shifts in judgments of IPs (62%) was in accordance with the FCE results about that activity. Direction of alternation was both in the direction of more as less physical work ability. It was concluded that FCE information does influence the judgment of IPs in the appraisal of disability claimants with MSD.

Chapter 7 describes the results of a descriptive study of the complementary value of FCE information for IPs, carried out with the aid of a questionnaire specially developed for this purpose. The first, and most important, question asked was whether the IP considered that FCE information was of complementary value for his assessment of physical work ability. The second question was split into two parts: a) whether the FCE information caused the IP to alter his assessment of the ability of the claimant examined to perform twelve work-related activities, and if so in what direction; and b) whether the IP considered that the FCE information had reinforced his opinion about the claimant's physical work ability. Finally, the IPs were asked whether they considered including FCE information in future disability claim assessment procedures, and if so for what groups of claimants. The minimum sample size needed to reject the hypothesis that IPs do not consider FCE information to be of complementary value for the judgment of physical work ability in disability claim assessments is 28 IPs. Of the 54 IPs who were prepared to participate in the study, 28 saw claimants who agreed to take part in the study and underwent FCE testing. Sixty-eight percent of these 28 IPs considered FCE information to be of complementary value. Since this exceeds the significance threshold of 66%, it is concluded that the IPs do consider FCE information to be of complementary value. Twenty-four IPs changed their assessment of the ability of the claimant to perform on one or more of the work-related activities considered after presentation of the FCE information. The number of changed judgements about the ability to perform the twelve activities was significantly higher in the group of IPs that did versus the

group that did not consider the FCE information of complementary value. Seventy-six percent of the IPs stated that the FCE information had confirmed their judgment. Finally, 15 of the 20 IPs who responded to the last question stated that they intended to include FCE information in future disability claim procedures. The claimant groups that were mentioned as likely to benefit from FCE testing were those with MSD (more specifically, whiplash, fibromyalgia and repetitive strain injury) and those with medically unexplained disorders. A strong relation exists between rating the FCE info as of complementary value and the intention of using FCE information in future disability claim assessments. The overall conclusion was that FCE information is of complementary value for the assessment of physical work ability by IPs in the context of disability claim procedures. From both studies can be concluded that EK FCE information has effect on the IP judgment and is considered to be of complementary value for the assessment of physical work ability of patients with MSDs.

Finally, Chapter 8 gives a general discussion, in which methodological considerations of the studies are taken into account. The overall conclusion is that FCE information is useful for IPs in their assessment of the physical work ability of disability benefit claimants. This leads to the suggestion that FCE information perhaps in combination with information from other sources can yield a broad, well-argued decision concerning the claimant's overall work ability. Directions for future research are identified, like a study about the criterion and construct validity of the EK FCE.

Samenvatting

Het beoordelen van het fysieke werkvermogen, de lichamelijke mogelijkheden die een patiënt heeft om te werken, van patiënten die een arbeidsongeschiktheidsuitkering hebben aangevraagd, is ingewikkeld en verzekeringsartsen die zijn aangesteld om dit te beoordelen hebben weinig instrumenten die hen daarbij kunnen helpen. Bovendien heeft het onderzoek naar de mate van arbeidsongeschiktheid te maken met het uitvoeren van een wet, de WAO of sinds 1 januari 2004 de WIA. Dit betekent dat een verzekeringsarts ook te maken krijgt met de wettelijke bepalingen en regels over de uitvoering van de wet. Op basis van deze wet heeft de patiënt een claim ten aanzien van het niet of niet volledig kunnen werken en verzoekt om een financiële compensatie daarvoor en de verzekeringsarts moet die claim beoordelen. Dit betekent dat de verzekeringsarts moet beoordelen wat het werkvermogen is. Dat beoordelen van het werkvermogen lijkt op het proces dat artsen volgen als ze de diagnose van een aandoening stellen. Het gaat alleen nu niet om de vraag welke diagnose de patiënt heeft maar wat zijn werkvermogen is. Kenmerkend voor zo'n diagnostisch proces is de onzekerheid over wat de juiste uitkomst is. Door informatie te verzamelen probeert de verzekeringsarts een oordeel te vormen over het werkvermogen. Belangrijk onderdeel van het werkvermogen is het fysieke werkvermogen. Om dit te kunnen beoordelen steunt de verzekeringsarts erg op de informatie die hij hierover van de patiënt krijgt. Een andere manier om het fysieke werkvermogen van patiënten te beoordelen is om niet alleen aan patiënten te vragen wat ze kunnen, maar door ze ook activiteiten te laten uitvoeren en vast te leggen in hoeverre dit mogelijk is. Dit is nu waar het bij de Functionele Capaciteit Evaluatie (FCE) methoden eigenlijk om gaat. FCE is een instrument dat het fysieke werkvermogen van patiënten vastlegt door te meten en te registreren hoe een patiënt die fysieke activiteiten uitvoert. Het gaat hierbij om activiteiten die in werk voorkomen, zoals staan, lopen, tillen, reiken, bukken, etc. Hoe betrouwbaar en valide FCE is, is in een beperkt aantal studies onderzocht en vrijwel altijd gebeurde dit bij revalidatie-patiënten of bij vragen over terugkeer naar werk. Dit proefschrift gaat om de vraag hoe nuttig FCE informatie is bij beoordelingen van de mate van arbeidsongeschiktheid. Dit leidt tot de volgende onderzoeksvraag:

- Wat is het nut van FCE informatie voor het oordeel van verzekeringsartsen over het fysieke werkvermogen in het kader van arbeidsongeschiktheidsbeoordelingen van werknemers met aandoeningen aan het bewegingsapparaat?

Om deze vraag te kunnen beantwoorden zijn zes studies verricht die beschreven zijn in de hoofdstukken twee tot en met zeven.

Hoofdstuk twee en drie zijn beide een systematisch onderzoek van de wetenschappelijke literatuur. In het eerste onderzoek (hoofdstuk 2) ging het om de vraag welke instrumenten (vragenlijsten en testen) gebruikt kunnen worden om te onderzoeken wat de fysieke mogelijkheden zijn van patiënten die een aandoening aan het bewegingsapparaat hebben. Voorwaarde om geselecteerd te worden was dat de vragenlijst of test gebruikt werd om activiteiten te onderzoeken in werk, algemeen dagelijkse levensverrichtingen of sport. Van de geselecteerde instrumenten is vervolgens onderzocht wat de betrouwbaarheid en validiteit ervan is. Het resultaat was 34 studies, waarvan vier vragenlijsten (de Oswestry Disability Questionnaire, de Pain Disability Rating Index, de Roland-Morris Disability Questionnaire, en de Upper Extremity Functional Scale) betrouwbaar en valide bleken te zijn. Geen van de testen die in de studies gebruikt werden, bleek zowel betrouwbaar als valide te zijn. Om een zo volledig mogelijk beeld te krijgen van het fysieke werkvermogen werd voorgesteld om een vragenlijst en een functionele test te combineren. Het tweede literatuuronderzoek (hoofdstuk 3) ging om de vraag wat er bekend is over de betrouwbaarheid en validiteit van een viertal FCE methoden. De vier FCE methoden waar het onderzoek op gericht was, waren: de Blankenship methode (BM), de Ergos Work simulator (EWS), de Ergo-Kit FCE (EK FCE) en de Isernhagen Work system (IWS). Er werden 77 studies geselecteerd die belangrijk zouden kunnen zijn voor verder onderzoek. Er bleven 12 studies over die gingen over betrouwbaarheid en validiteit van FCE methoden. Het bleek dat vooral de IWS op een aantal aspecten goede uitkomsten op betrouwbaarheid en validiteit had. Er werden geen studies gevonden die de betrouwbaarheid van de BM, EWS en EK FCE hadden onderzocht. Dit betekent dat er meer studies nodig zijn om de betrouwbaarheid en validiteit van FCE methoden aan te tonen en dan vooral van de BM, EWS en EK FCE.

In hoofdstuk 4 wordt een onderzoek beschreven dat gaat over de betrouwbaarheid en overeenkomst tussen testleiders van vijf tiltesten van de EK FCE bij mensen met lage rugklachten. Vierentwintig patiënten deden mee aan het onderzoek en zij kwamen uit een aantal verschillende fysiotherapiepraktijken. De patiënten werden getest door twee testleiders op twee verschillende momenten met een tussenpoos van drie dagen. De resultaten lieten hoge uitkomsten zien op het niveau van betrouwbaarheid van de tiltesten. Ook de overeenkomst tussen de testleiders ten aanzien van de resultaten van de tiltesten was goed.

Dus kon uit het onderzoek geconcludeerd worden dat bij patiënten met lage rugklachten de EK tiltesten betrouwbaar zijn en de overeenkomst tussen testleiders goed is.

In hoofdstuk 5 worden de resultaten beschreven van een interview onderzoek naar hoe nuttig FCE als instrument bij re-integratie- en claimbeoordelingsprocedures is om het fysieke werkvermogen vast te stellen. Eenentwintig case managers die zich bezig houden met re-integratie en 29 claimbeoordelingsexperts werden telefonisch geïnterviewd aan de hand van een zelf ontwikkelde vragenlijst. Er werden vragen gesteld over hoe nuttig de ondervraagden FCE voor hun werk vonden en op grond van welke argumenten. Ook werd ze gevraagd of er misschien randvoorwaarden bestonden als het gaat om FCE testen te laten doen. De deelnemers moesten ervaring hebben met FCE informatie in hun werk. Re-integratie case-managers waardeerden het nut van FCE met een gemiddelde van 6,5 (SD 1,5) op een schaal van 0-10. Claimbeoordelingsexperts waardeerden het nut van FCE gemiddeld lager en de spreiding was groter met een gemiddelde van 4,8 (SD 2,2). Argumenten om FCE nuttig te vinden waren: het versterkt het eigen oordeel en het is een objectieve manier van meten. Argumenten tegen het nut van FCE waren: het ontbreken van nieuwe informatie en juist het gebrek aan objectiviteit van de FCE meting. Redenen om een FCE onderzoek te laten doen waren: aandoeningen aan het bewegingsapparaat, een positieve visie van de patiënt over zijn mogelijkheden om te werken en de beschikbaarheid van een concrete baan. Redenen om geen FCE te laten doen waren: moeilijk objectiveerbare aandoeningen, een negatieve visie van de patiënt over zijn arbeidsmogelijkheden en het bestaan van een geschil of een juridische procedure.

Hoofdstuk zes en zeven hebben betrekking op dezelfde verzekeringsartsen en patiënten met aandoeningen aan het bewegingsapparaat. In het eerste van de twee onderzoeken (hoofdstuk 6) worden de resultaten beschreven van een studie onder verzekeringsartsen. Het doel was te onderzoeken of FCE informatie effect heeft op het oordeel van verzekeringsartsen over het fysieke werkvermogen van patiënten die een aandoening hebben aan het bewegingsapparaat. Van de 100 verzekeringsartsen die willekeurig geselecteerd werden uit het UWV bestand van verzekeringsartsen die claimbeoordelingen doen, deden uiteindelijk 54 mee aan het onderzoek. De opzet was om van iedere verzekeringsarts twee patiënten met een aandoening aan het bewegingsapparaat te laten meedoen aan het onderzoek. De eerste patiënt kreeg een FCE onderzoek en de tweede fungeerde als controlepatiënt. De verzekeringsarts scoorde het fysieke werkvermogen van beide patiënten op een lijst met 12 activiteiten na

afloop van de gewone arbeids- ongeschiktheidsbeoordeling. De verzekeringsarts gaf aan wat het vermogen van de patiënt was om bepaalde activiteiten uit te voeren ten opzichte van de situatie voordat de patiënt de aandoening kreeg. De grenzen waarbinnen de verzekeringsarts kon scoren waren als volgt bepaald: score 0 % betekende dat de mogelijkheid voor de patiënt om die activiteit uit te voeren volledig onmogelijk was ten opzichte van de situatie voor het optreden van de aandoening. Score 100% betekende dat de mogelijkheid van de patiënt om die activiteit uit te voeren nog even groot was als voor het optreden van de aandoening. Het rapport van het FCE onderzoek werd bij het dossier van de patiënt gevoegd. Hierna werd aan de verzekeringsarts gevraagd het fysieke werkvermogen van de beide patiënten opnieuw te scoren voor dezelfde 12 activiteiten maar nu aan de hand van de beide patiëntdossiers. Bij de ene patiënt die het FCE onderzoek had ondergaan, was naast het patiëntendossier ook het FCE rapport beschikbaar. Bij de andere patiënt was alleen het patiëntendossier beschikbaar. In totaal deden 54 patiënten van 27 verzekeringsartsen mee aan de studie. Bij de helft van de patiënten werd een FCE onderzoek uitgevoerd. Het aantal keren dat de verzekeringsarts zijn oordeel veranderde tussen de eerste en tweede beoordeling over iedere activiteit, werd vastgelegd. Na toetsing bleek dat verzekeringsartsen hun oordeel vaker veranderden bij de groep van patiënten bij wie ze over FCE informatie beschikten dan bij de groep van patiënten bij wie deze informatie niet aanwezig was. In een meerderheid van de gevallen (62%) gingen de veranderingen in oordeel van de verzekeringsartsen dezelfde kant op als het resultaat van het FCE onderzoek. Veranderingen gingen zowel in de richting van minder als van meer mogelijkheden om die activiteit uit te voeren. De conclusie is dat FCE informatie invloed heeft op het oordeel van verzekeringsartsen als het gaat om beoordelingen van de mate van fysiek werkvermogen van patiënten met aandoeningen aan het bewegingsapparaat. In het tweede onderzoek (hoofdstuk7) bij dezelfde groep van verzekeringsartsen worden de resultaten beschreven van een beschrijvend onderzoek naar wat verzekeringsartsen vinden van de toegevoegde waarde van FCE informatie voor hun oordeel over het fysieke werkvermogen. Drie vragen werden aan de verzekeringsartsen voorgelegd na afloop van het FCE onderzoek en het aanbieden van het FCE rapport. De eerste en belangrijkste vraag was of de verzekeringsartsen meenden dat FCE informatie een toegevoegde waarde heeft voor hun oordeel over het fysieke werkvermogen. De tweede vraag was of deze informatie reden was om het oordeel over het fysieke werkvermogen te veranderen, en zo ja, welke kant op. Het tweede deel van die vraag was of de FCE informatie hun oordeel over het fysieke werkvermogen had versterkt. Tenslotte werd de verzekeringsartsen gevraagd of zij er over dachten om FCE informatie in toekomstige beoordelingen van arbeidsongeschiktheid

opnieuw te gaan gebruiken en zo ja, bij welke patiënten zij dat dan zouden doen. Achtentwintig verzekeringsartsen en patiënten waren nodig en hebben ook meegedaan aan het onderzoek. Uit de resultaten bleek dat de meerderheid van de verzekeringsartsen (68%) van mening was dat FCE een toegevoegde waarde heeft. Op vier na veranderden alle verzekeringsartsen hun oordeel op basis van de informatie op één of meer activiteiten. Een ruime meerderheid van de verzekeringsartsen (76%) gaf aan dat FCE informatie hun oordeel over het fysieke werkvermogen had versterkt. Vijftien van de 20 verzekeringsartsen die deze vraag beantwoordden, gaven aan dat zij ook in toekomst FCE informatie zouden willen betrekken bij claimbeoordelingen. De patiëntengroepen die zij hierbij vooral op het oog hadden, waren: patiënten met aandoeningen aan het bewegingsapparaat en meer specifiek whiplash, fibromyalgie, repetitive strain injury (RSI) en de groep van medisch onverklaarde aandoeningen. Verzekeringsartsen die vonden dat FCE een toegevoegde waarde heeft, gaven aan dat ze ook in de toekomst FCE informatie bij arbeidsongeschiktheidsbeoordelingen willen betrekken. De conclusie is dat FCE informatie toegevoegde waarde heeft voor het oordeel over het fysieke werkvermogen van verzekeringsartsen bij de beoordeling van arbeidsongeschiktheid.

De discussie in hoofdstuk 8 over de zes hoofdstukken beschrijft overwegingen die te maken hebben met de opzet en uitvoering van de verschillende onderzoeken en gaat in op de betekenis van de gevonden resultaten. Het antwoord op de hoofdvraag is dat FCE informatie nuttig is voor het oordeel van verzekeringsartsen over het fysieke werkvermogen in het kader van arbeidsongeschiktheidsbeoordelingen. Dit betekent dat FCE informatie een rol kan spelen in het proces van arbeidsongeschiktheidsbeoordelingen. Het antwoord op de hoofdvraag roept ook nieuwe vragen op die te maken hebben met de vraag welke patiënten nu het meest geschikt zijn voor een FCE onderzoek. Ook de vraag naar het nut van FCE informatie als rekening wordt gehouden met de bijzondere omstandigheden rondom de beoordeling, is met dit onderzoek niet beantwoord. Het gaat dan om omstandigheden waarin sprake is van letselschade, bezwaar- en beroepsprocedures tegen beslissingen over de mate van arbeidsongeschiktheid, en dergelijke. De aanbevelingen uit dit onderzoek richten zich ook op het meenemen van meer bronnen van informatie zoals vragenlijsten en testen bij de beoordeling van het werkvermogen in het kader van arbeidsongeschiktheidsbeoordelingen door verzekeringsartsen. Verzekeringsartsen zullen dan wel geschoold moeten worden zodat zij de informatie die dan beschikbaar komt op een goede manier leren gebruiken bij hun oordeel.

Desondanks blijft het vaststellen van het werkvermogen een moeilijk proces. De onzekerheid over de juiste uitkomst wordt met dit onderzoek niet weggenomen. De resultaten van deze studies tonen wel aan dat FCE-testen een toegevoegde waarde hebben voor de beoordeling van het fysieke werkvermogen bij patiënten met aandoeningen aan het bewegingsapparaat door verzekeringsartsen in een beoordeling van de mate van arbeidsongeschiktheid.

Dankwoord

‘Last but certainly not least ……’
Iedereen die heeft meegewerkt aan het tot stand komen van dit onderzoek wil ik graag hartelijk bedanken. In de afgelopen 4 ½ jaar zijn er veel mensen geweest die allemaal op hun manier hebben bijgedragen aan dit proefschrift en zonder die bijdragen zou het nooit gelukt zijn. Iedereen die op welke manier dan ook een bijdrage heeft geleverd aan deze studie, hartelijk dank voor jullie inbreng en inzet. Het voert te ver om iedereen persoonlijk te bedanken maar een enkeling wil ik hier wel noemen.

In de eerste plaats wil ik mijn promotor Monique Frings-Dresen noemen. Voor jou was het een hele opgave een kandidaat te begeleiden die qua leeftijd dan misschien wel je gelijke was maar zich wat wetenschappelijk denken betreft nog in de kinderschoenen stond. Dat vraagt om geduld en durf en over beide eigenschappen beschik je in ruime mate. Het vakgebied, de verzekeringsgeneeskunde was je vreemd en in de afgelopen jaren heeft de wetenschappelijke naïviteit van mijn kant gecombineerd met het merkwaardige werkterrein dat verzekeringsgeneeskunde heet, regelmatig tot flinke discussies geleid. Maar als de stofwolken dan weer waren neergedaald, bleek telkens weer dat wetenschappelijke expertise meer waard is dan praktisch empirisme.

Paul Kuijer, co-promotor, als geen ander ben je van begin af aan blijven geloven in de goede afloop. Met je eeuwig optimisme wist je altijd, ook op momenten dat het even niet meer ging, een positieve wending aan het geheel te geven. Met je beruchte pennetje had je altijd wel wat op te merken maar aarzelde je niet om ook te schrijven dat je iets goed vond. Niets stimuleert meer dan te horen of te lezen dat het goed is.
En niet alleen op het Coronel kon ik voor vragen bij je terecht. Ook daarbuiten was er de gastvrijheid, het begeleiden hield niet op bij de drempel van het AMC.

Judith Sluiter, co-promotor, jij kwam wat later bij het begeleidingsteam. Je kennis en creativiteit hebben een grote impuls gegeven aan het hele project. Met humor en deskundigheid had je een heel eigen inbreng in de wekelijkse besprekingen. Met je altijd positief gestemde kritische opmerkingen heb je er mede voor gezorgd dat het werk de toets van voldoende wetenschappelijk niveau kan doorstaan.

Vincent, als kamergenoot en mede FCE- onderzoeker, heb je meer dan wie ook van dichtbij kunnen meemaken hoe een promotietraject er uit ziet. Altijd bereid tot het doen van niet de

leukste klussen heb je een belangrijke bijdrage gegeven aan het project. Vooral de hulp bij de voor mij zo lastige statistiek was heel waardevol.

De promotiecommissie bedank ik voor de tijd en aandacht die zij hebben besteed aan het proefschrift.

Frans Slebus, heel blij ben ik dat je bereid bent paranimf te zijn. Je was steun en toeverlaat op de momenten dat het even niet meer ging en die momenten waren er regelmatig. De wandelingetjes tussen de middag in weer en wind zorgden voor nieuwe inspiratie om door te gaan. Heel blij ben ik daarom dat ik ook bij de laatste loodjes op je kan terugvallen.
Hoewel je daar regelmatig je twijfel over uit, weet ik het zeker: binnen een jaar sta jij hier ook!

Jan Willem van Zadelhoff, blij ben ik dat jij ook paranimf wil zijn. Jij bent de onontbeerlijke link met UWV. In de achterliggende jaren heb je mij voortdurend gesteund met je belangstellende vragen. Verder zorg je ervoor dat ik op de hoogte blijf van de roerige ontwikkelingen binnen UWV in het algemeen en UWV Hengelo in het bijzonder.

Johan Oosterloo, zonder jou zou dit hele project niet mogelijk geweest zijn. Als jij mij indertijd niet geïnspireerd had, was het er vast niet van gekomen. Jij hebt ook als één van de eersten ingezien dat academisering van de verzekeringsgeneeskunde nodig is. Je hebt er voor gezorgd dat er een financieel draagvlak kwam waarmee het voor mij en anderen mogelijk werd om aan een promotietraject te beginnen.
Dank geldt daarom ook het UWV dat het belang van dit project maar ook van al die andere projecten binnen het Kenniscentrum Verzekeringsgeneeskunde heeft ingezien en ondersteunt.

De collega's in Hengelo wil ik danken voor hun belangstelling door de jaren heen. Even wat overleggen of even wat regelen ging niet want dan was ik weer in Amsterdam. Maar zonder daar ooit een opmerking over te maken, werd dat gewoon geaccepteerd. Tekenend voor de flexibiliteit en collegialiteit die het kenmerk zijn van UWV Hengelo. Ook het management van UWV Hengelo wil ik bedanken voor hun bereidheid om mee te denken en mee te werken zodat het project goed afgerond kon worden.

De 'Coronellers' wil ik danken voor het zonder meer opnemen van mij in de kring van AIO's.

Terwijl jullie allemaal nog min of meer aan het begin van jullie beroepscarrière staan, was daar zo maar iemand die al jaren werkt en zelfs al opa is. Dat moet toch wel wat vreemd geweest zijn. Jullie hebben altijd van harte meegeleefd met de hoogte- en dieptepunten die een traject als dit nu eenmaal kenmerken. Dank daarvoor.

Alle verzekeringsartsen, patiënten maar ook de Ergo Kit medewerkers wil ik bedanken voor het meewerken aan dit onderzoek. Het is een beetje cliché maar daarom niet minder waar dat zonder jullie het onderzoek nooit gedaan had kunnen worden. De Ergo Kit medewerkers en dan vooral Jan Plat, Niels Geise en Laurens van der Kraats wil ik hartelijk danken voor de medewerking die het mogelijk maakte bij de FCE testen gebruik te maken van de Ergo Kit FCE.

Veel mensen hebben in de loop van de jaren meegeleefd en meegedacht met het project dat nu tot stand gekomen is maar twee wil ik toch graag noemen. Jos Harsta, bij het eerste en ook bij het laatste artikel waren je kennis en deskundigheid onontbeerlijk om mij met de zo lastige grammatica van het Engels te helpen. Oom Jan Bierling, hartelijk dank voor het helpen met de taalkundige voetangels en klemmen bij het schrijven van de Nederlandse samenvatting.

Voor jullie, mijn kinderen is het geen gemakkelijke tijd geweest. Er was vaak weinig tijd voor jullie maar hopelijk gaat dat in de komende tijd veranderen.

Petra, er zijn bijna geen woorden voor om aan te geven hoe belangrijk jouw steun was en is. Allerlei veranderingen waren het gevolg van mijn keus om dit traject aan te gaan. Ik noem maar het veranderen van de goede Hollandse gewoonte dat om 6 uur de aardappels op tafel staan naar de meer Franse manier van leven waarbij 8 uur ook nog een keurige tijd is om te dineren. Voor het stilzwijgende opvangen en wachten, zonder daarover ooit te klagen of te mopperen, ook als het weer eens helemaal niet meer ging, daarvoor wil ik je danken. Ik zou nog zoveel meer willen zeggen maar jij vindt dat dit niet in zo'n boekje hoort en daar heb je gelijk in.

Het boek is nu af en dat is goed.

Publications

The following parts of this thesis have been published:

Wind H, Gouttebarge V, Kuijer PPFM, Frings-Dresen MHW. Assessment of functional capacity of the musculoskeletal system in the context of work, daily living and sport: a systematic review. J Occup Rehabil 2005; 15 (2): 253-272 (Chapter 2)

Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. 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 2004; 77: 527- 537 (Chapter 3)

Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. Reliability and agreement of 5 Ergo-Kit functional capacity evaluation lifting tests in subjects with low back pain. Arch Phys Med Rehabil. 2006; 87(10): 1365-1370 (Chapter 4)

Gouttebarge V, Wind H, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. Betrouwbaarheid en overeenkomst van vijf tiltests bij mensen met lage rugklachten. Tijdschr Bedrijfs Verzekeringsgeneesk. 2007; 4: 156-161 (Chapter 4)

Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. The utility of functional capacity evaluation: the opinion of physicians and other experts in the field of return to work and disability claims. Int Arch Occup Environ Health. 2006 Jun;79(6):528-34 (Chapter 5)

Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. Het nut van functionele capaciteit evaluatie: de visie van experts. Tijdschr Bedr Verzekeringsgeneesk. 2005; 13 (10): 300-305 (Hoofdstuk 5)