Quality of functional capacity evaluation tests: a clinimetric approach
Gouttebarge, V.

Citation for published version (APA):

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

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

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Chapter 1

General introduction
Chapter 1  General introduction

Background

Why do satellites span the sky? Why do countries send spies into other countries? Why do professional sport teams study their opponents' games? The answer is because people today are not reluctant to gather information. This process empowers the development of knowledge and reduces, to some extent, uncertainties in judgement and decision making processes. Consequently, new instruments and tests are constantly being developed and fashioned to provide complementary meaningful information to the search for information. Whether a test’s outcome can provide trustworthy judgements or decisions depends particularly on the measurement quality of the test (1).

Whereas biometrics and psychometrics concern the sciences of measuring and analysing biological and psychological phenomena, respectively, clinimetrics refers analogously to the development of methodological and statistical methods applicable in clinical medicine (2,3). Clinimetrics deals with the design, administration and interpretation of tests meant to functionally and accurately measure clinical and epidemiological variables, such as symptoms, diagnosis, progression of illness or problems of functional capacity in daily living and work (2,3). Assurance of accurate measurement is required before using tests in any given context. Clinimetric properties indicating that the test is reliable and valid should be considered as fundamental for determining the measurement quality of any test (4).

In light of the financial burden of musculoskeletal disorders seen worldwide, properly and accurately assessing physical work-ability is crucial, especially in rehabilitation, return to work and work disability contexts. However, uncertain assumptions are made for evaluation, prediction, judgement or decision making related to physical work ability (5). Professionals, working either in rehabilitation, occupational or insurance medicine, tend to rely on information provided by the patient’s recall and the physical examination and draw inferences to estimate the physical work-ability of a patient (6). Complementary information, provided by relevant performance-based tests such as Functional Capacity Evaluation (FCE) tests, could allow professionals to base their
physical work-ability assessment on other than intuitive principles or presumptive facts, but rather on scientific and accurate evidence if, and only if the FCE tests concerned are reliable and valid.

Clinimetrics

Derived from the field of psychometrics, the term ‘clinimetrics’ was introduced in the early 1980s by Feinstein, and it is defined as “the domain concerned with indexes, rating scales and other expressions that are used to describe and measure symptoms, physical signs, and other distinctly clinical phenomena in clinical medicine” (4,7). In recent years, many authors have focused and debated on the similarities and differences between clinimetrics and psychometrics. Streiner states that clinimetrics represents a fraction of what psychometrics already includes and suggests that the distinction is unnecessary (8). Several years ago, de Vet emphasised the substantial overlap between clinimetrics, biometrics (science measuring human biological and physical characteristics) and psychometrics, but emphasised that clinimetrics is becoming a fundamental methodological discipline within modern medicine (2,3). Indeed, psychometrics has been developed outside the clinical field, mainly in the educational and social areas, and it is a long-standing discipline of the methodology for measuring psychologic phenomena or educational achievements (9,10). Therefore, rating scales based on self-report by the patients have been developed and used in order to quantify unobservable psychological phenomena, such as anxiety or stress (10,11). Analogously, clinimetrics refers generally to the development of methodological and statistical methods applicable in clinical medicine in order to assign numbers or scores to observable clinical events (2,3,9). Aside from self-report questionnaires, clinimetrics also deals with the development, design, administration and interpretation of performance-based instruments or tests that are assessed by clinicians or evaluators (9).
Today, clinimetrics focuses on the measurement quality provided by existing clinical instruments, and quality assessment is a fundamental and necessary process. In fact, information provided by any clinical instrument cannot be trusted and licitly used in any judgement and decision making process if the measurement quality (i.e., clinimetrics) has not been positively evaluated. In the development process of any instrument, measurement quality is a topic of concern and is reviewed through a series of steps, including generation of proper items or subtests, selection of proper outcome measures, standardization of material and assessment procedures, and attention to interpretability of test results. However, even if the process of an instrument development deals, to some extent, with minimizing measurement error, clinimetrics remains an imperative process (Figure 1). Following development (Figure 1, nr 1 & 2), any clinical instrument or test cannot be applied or used for a given purpose for any given context (Figure 1, nr 4) until it has passed the clinimetrics assessment (Figure 1, nr 3). If an instrument or test fails upon clinimetrics assessment (Figure 1, nr 5), it cannot be licitly implemented, and, consequently, its development should be reconsidered. Therefore, identifying potential sources of measurement error could eventually lead to the correct adjustments and modifications to improve measurement quality.

Figure 1: From development to use of clinical instruments in a certain context
Among the different measurement theories originating from psychometrics, classical test theory (CTT) has been by far the most influential test theory applied to clinimetrics to assess the measurement quality of clinical instruments (1,11,12). Similar to true score theory, CTT is a simple yet powerful model for the assessment of measurement quality, and it relies on the decomposition of observed scores to true and error scores (1,11,12). The measurement of any clinical phenomenon, such as blood pressure, maximal oxygen uptake or lifting capacity, carries some amount of error that can be divided into random error and systematic error (12). Random error is present in all measurements and is caused by factors that randomly and unpredictably affect the measurement of a phenomenon (1,12). For instance, mood and motivation could be factors that randomly increase or decrease a person’s performance. Analogously, systematic error is caused by factors that systematically affect the measurement of a phenomenon, and this error will consistently either increase or decrease the observed score (1,12). For instance, a personal weight scale could deliver measurements that are invalid but reproducible, such as always registering a value that is two kilograms too high because of an improper calibration. Measurement errors, both random and systematic, are the core of clinimetrics and are traditionally associated with the two key concepts of reliability and validity (12). In clinical research, both clinimetric properties are relevant for performance-based instruments, and reliability is the first clinimetric property that must be assessed because no test can be valid without being reliable (1,11).

**Reliability**

Reliability assessment is the first step in the evaluation of the clinimetric properties of an instrument; a test cannot be valid if it is not reliable. Reliability refers to the consistency of a measure, and it reflects the amount of random error involved in any measurement (1,11). In the literature, reliability is often used interchangeably with the terms reproducibility, repeatability, consistency, agreement and stability (13). Recently, de Vet advocated that reproducibility is the proper term to use in clinical research, making the distinction between two aspects that are essential for clinical interpretation: reliability and agreement (2,14).
Reliability refers to the test’s ability to distinguish individuals from each other despite measurement errors (2,13,15). The terms intra- and interrater reliability apply when the repeated measurements are assessed within or between raters, respectively (13). Agreement concerns the absolute measurement error in that it evaluates how close scores are during repeated measurements. In agreement assessment, the measurement error is established within raters when the repeated measurements are assessed by the same rater, or between raters when assessed by different raters (2,15). This quantification of measurement error permits cautious assumptions about sensitivity to change, or responsiveness; a large measurement error might not permit separation of real change from changes due to measurement error, making it useless for detecting any relevant clinical changes in time. With regard to clinical medicine, de Vet acknowledges agreement as a particularly important parameter because a test is often used in clinical medicine to assess changes over time (14,15). Furthermore, agreement, by comparison with reliability, has the advantage of being expressed on its own measurement scale, which makes it easier to interpret (14,15). However, in case of unfamiliarity with test scores, for instance, in assessing a newly developed instrument, clinical interpretation of agreement is difficult to assess because a given change in test scores cannot be clearly assigned to real clinical change or measurement error, while reliability parameters can be interpreted fairly well (14,15).

For both agreement and reliability, the phenomenon that is supposed to be measured presumably does not change (steady-state) between the repeated measurements. Therefore, from a methodological and procedural perspective, the choice of an optimal time interval is then an essential aspect in reproducibility assessment, especially when assessing reproducibility in a population with a medical condition (either disorder or complaint). In medical conditions, the time interval chosen should take into account the nature, course and temporal stability of the disorder or complaint. Generally, time intervals in reproducibility studies are accepted to vary from a few days to two weeks between repeated measurements (10,11,16).

Finally, it is essential to indicate that reproducibility is a property of the test scores rather than the test itself, and it is thus said to be sample dependent (11). Since
particular tests are intended to be applied to certain populations, reproducibility studies should assess populations that are relevant for future use and implementation (once evidence of reproducibility and validity has been established) of the concerned test. The test in question is then shown to be reproducible only within the population that has been used in the reproducibility study.

**Validity**

Validity assessment is a major issue in the assessment of the quality of a test. This assessment is not a simple matter because there are several definitions in literature and textbooks, and there are many methodological and procedural approaches. While reproducibility refers to the quality of a test’s scores and relates to random error, validity reflects the quality of the interpretation of test scores and relates them to systematic error. In other words, validity shows that a test measures what it intends to measure (1,10-12). In clinical research, current and accepted validity concepts include content, construct and criterion-related validity, the last two being the most relevant for performance-based tests (11).

Content validity can be claimed when a test logically and obviously measures what it purposes to measure (1,17). Content validity does not need statistical assistance to be established because the relationship between the phenomenon being measured and the test score(s) is determined by a panel of experts or researchers (17). In a clinical setting, because of the subjectivity involved in the assessment of content validity, researchers prefer to seek construct and criterion-related validity evidence, which can be statistically supported. Also, within psychometrics, content validity appears as a ‘category of construct validity’ that is not always assessed for (psychological) instruments (18).

Construct validity refers to the degree to which a test measures a hypothetical, non-observable construct, and this validity can be established by relating the test to outcomes of other instruments (1,11). Construct validity is the broadest type of validity and can be subdivided into convergent, divergent and discriminative validity (1,11,17).
Convergent validity measures the degree to which the evaluated test is associated with another test that is believed to reflect the same underlying phenomenon. Divergent validity, also referred to as discriminant validity, measures how well the evaluated test is associated with another test that is believed to reflect a different underlying phenomenon \((11,17)\). Discriminative validity, not to be confused with discriminant validity, is another common way to support construct validity and is provided when the evaluated test discriminates between groups that are known and expected to be different from each other \((1,17)\). This method is also referred as the Known Groups Method \((1)\).

Criterion-related validity can be subdivided into concurrent and predictive validity and is the most practical, powerful kind of validity, and the one most needed by clinicians. Criterion-related validity describes how the evaluated test relates to an existing highly valued test, called a gold standard (showed to be reproducible and valid) that measures the same concrete phenomenon \((11,17)\). Concurrent validity examines the relation between both tests concurrently, while predictive validity examines the same relation, but the gold standard is measured at a later time \((11,17)\). The most difficult aspect of establishing criterion-related validity in clinical practice is that often no gold standard exists \((19)\). When no gold standard is available, the common alternative is to use an accepted and well-grounded reference test (also referred as silver standard) to relate to the evaluated test \((20,21)\).

The validation of clinical performance-based tests, as this thesis will deal with, is not a straightforward process. After evaluating the basic property of reproducibility in a relevant population, the validation process of performance-based tests requires gathering evidence on whether the test results are meaningful in a given context. Therefore, a series of studies assessing different aspects of validity could be used, in which evidence of construct and criterion-related validity is preferred. With regard to criterion-related validity, there are few unequivocal instruments that can act as a gold standard for performance-based tests, so an adequate alternative or reference test must be found.
Physical work-ability and Functional Capacity Evaluation

Physical health, either for work, sport or activities of daily living, is an important concern for the quality of life. Proper functioning of the musculoskeletal system is an important determinant of physical health. Unfortunately, the musculoskeletal system does not always cooperate. For example, in many populations, either musculoskeletal complaints (MSCs) or musculoskeletal disorders (MSDs) cause health problems and pervade every area of life, especially work (22-25). Consequently, physical impairments often emerge while working, thus hindering physical work-ability. Work-related physical impairments are, therefore, a significant health care problem with regard to both incidence and costs.

Indeed, MSDs have been identified as the most common cause of occupational disease and the primary reason for long-term absence and related work disability (26). In addition, 60 to 75% of the people suffering from MSCs in Europe have a decreased ability to perform physical activities of daily living (27), which places a tremendous financial burden on health services. In the United Kingdom, the direct cost of one workday lost resulting from absence due to MSCs has been estimated to be 450 Euro (28). The total direct cost for health services that result from MSDs was reported to be 0.7% of the gross national product in the Netherlands, 1.0% in Canada, and 1.2% in the USA (29,30). Furthermore, aside from psychological disorders, MSDs are the most common cause for work disability and sick leave in the Dutch working population, representing an expenditure of more than 10 million Euro over a year (31). In order to reduce the financial burden of MSD management, it is imperative that professionals concerned with rehabilitation and disability be able to adequately assess physical work-ability. However, these professionals possess few instruments to assess physical work-ability, and they often rely on their own experience and on information provided by the patient’s anamnesis, the physical examination, and information from the health sector (6,32). Consequently, the use of complementary information provided by a relevant performance-based test specifically developed to assess physical work-ability would improve judgement and the decision making process.
Chapter 1  General introduction

Functional Capacity Evaluation (FCE) tests are performance-based assessments designed to assess the current physical work-ability of individuals suffering from MSCs or MSDs (33-35). FCEs are believed to have originated in the 1970s from American therapists who assembled and compiled existing and self-developed tests into a battery of tests that simulate work and assess the patient’s ability to work (36). FCEs rely on the results of a battery of standardized functional tests that reflect work-related activities, such as sitting, standing, walking, bending, reaching, hand and finger dexterity, lifting and carrying (33-35). Based on observation and testing criteria, FCE tests evaluate the performance of individuals in tasks of both short and long duration (37,38). FCE tests report several factors, including blood pressure, heart rate, liftable load, working height, working distance, manipulation velocity, coordination, degree of pain, and fatigue (37,38). FCE tests are generally administered by certified physical or occupational therapists specifically educated and regularly trained in adequate administration. The results of FCE tests can be collected on-line, in the case of computerized FCEs, or off-line, where data are filled in on an appropriate form by the evaluator. Standard FCE protocols are generally time-consuming. It can take between four hours and two days to go through the entire FCE protocol, depending on the type of FCE, the testing procedures and criteria, the evaluator’s skills and the individual’s abilities. Ease of administration and time cost make FCEs relatively impractical, practicality being a critical issue when it comes to FCE assessment.

FCEs are generally applied to different (para)medical contexts for several purposes (39-42). In rehabilitation medicine, FCE tests can be used as an initial medical or clinical evaluation following injury to assess a patient’s physical abilities. They can also be used as an evaluative and training instrument to determine physical and functional improvement through a rehabilitation program. In occupational medicine, FCE tests can give relevant support to occupational physicians to formulate appropriate and realistic expectations in terms of return to work planning. Here, FCE tests can also propose adequate and hopefully effective rehabilitation or treatment programs that could successfully lead to full or modified return to work. In insurance medicine, insurance physicians can use FCE tests to determine the degree of physical (dis)ability for work disability compensation purposes and to provide meaningful information to justify work
disability benefits. From the patient’s perspective, an FCE can provide a patient with valuable feedback about his or her own physical abilities because the FCE focuses on a person’s abilities and physical capacities given a medical condition, disability or impairments. Furthermore, the patient can personally experience what he or she is capable of doing despite the medical condition.

In the Netherlands, four major FCEs are being profiled on the Dutch market. One of these tests, the Ergo-Kit (EK) FCE method, is particularly widespread (43). The EK FCE is based on 55 tests, including 15 ‘physical agility’ tests that are principally used by physical and rehabilitation therapists and applied in physiotherapy and rehabilitation centers as an evaluative and training instrument. In work disability context, insurance physicians recently indicated that information provided by different EK FCE tests have complementary value in judging physical work-ability (44,45). Furthermore, occupational and insurance physicians, working either in return to work or work disability context, indicated a positive view of the utility of complementary information provided by the EK FCE tests for the assessment of the physical work-ability of a patient. However, this view applies only under the assumption that the measurement quality of the EK FCE tests has been assessed and found to be positive (46).

Clinimetrics of the Ergo-Kit FCE tests

Despite the providers statement that FCEs are reliable and valid, Innes and Strakker claimed that there is too little scientific information on the measurement quality, or clinimetrics, of most FCEs (13,17). In addition, Gardener and colleagues suggested that the lack of documented reliability and validity further diminishes confidence in any approach to FCEs (47,48). Consequently, the search for evidence on measurement quality of EK FCE tests either in the literature or through empirical study is imperative before implementing it in the context of rehabilitation or occupational or insurance medicine. With regard to the theoretical and methodological considerations of clinimetrics described earlier in this chapter, the measurement quality of EK FCE tests
should be represented in a two-step model assessing reproducibility and validity successively (Figure 2).

Reducing, or even eliminating, error in any measurement or observation, including EK FCE test scores, remains the main objective of clinicians. Therefore, both clinimetric properties can be approached according to the CTT to relate variance in EK FCE test scores to random or systematic error. Assessing reproducibility (Figure 2, nr 1) of EK FCE tests could allow quantification of random error, while the evaluation of validity (Figure 2, nr 2) could identify potential sources of systematic error. Only once clinimetrics has been positively evaluated information provided by EK FCE tests can be trusted and eventually incorporated into the assessment of physical work-ability in a specific and defined context. The seek of clinimetric evidence for EK FCE tests, as illustrated through a two-step model (Figure 2), will be covered in this thesis.
Objective of this thesis and research questions

As illustrated in figure 2, the main objective of this thesis is to find scientific evidence for the quality of FCE tests, focussing on whether FCE tests give reproducible outcomes in repeated measurements and whether the outcomes of FCE tests are valid for the assessment of physical work-ability in a return to work context.

With regard to the reproducibility of FCE tests, the following three research questions will be answered in this thesis:

1. What is known in the international literature about the reliability of four FCE methods available in the Netherlands?
2. How reproducible are EK FCE tests in subjects without musculoskeletal complaints?
3. How reproducible are EK FCE tests in subjects suffering from musculoskeletal complaints?

With regard to the validity of FCE tests, the following three research questions will be answered in this thesis:

4. What is known in the international literature about the validity of four FCE methods available in the Netherlands?
5. What is the construct validity of EK FCE tests in employees on sick leave due to MSDs?
6. What is the criterion-related validity of EK FCE tests in employees on sick leave due to MSDs?

Due to the uncertainty of the clinimetrics of FCE tests, no specific hypotheses can be formulated on whether FCE tests provide valid or non-valid measurements for the assessment of physical work-ability.
Outline of this thesis

Chapter 2 provides a systematic literature review of the clinimetric properties of the four FCE methods that are profiled on the Dutch market, including the EK FCE. Chapter 3 describes a study evaluating the intra- and interrater reliability of functional tests from the EK FCE method in adults without musculoskeletal complaints. Chapter 4 describes a study evaluating the reproducibility (agreement and reliability) between raters of five EK FCE lifting tests in subjects suffering from low back pain. The measurement error of these functional tests is quantified to demonstrate the sensitivity to change within the study population. Chapter 5 describes a study evaluating the construct validity of five EK FCE lifting tests (two isometric and three dynamic) in construction workers on sick leave due to MSD. Discriminative validity is assessed using the Known Groups Method, while convergent validity is established by studying the relationship between these functional tests and pain intensity and disability. In chapter 6, criterion-related validity is established by evaluating the concurrent and predictive validity between five EK FCE lifting tests (two isometric and three dynamic) and a reference test, an instrument that is used and accepted among the occupational health services for workers in the construction industry on sick leave due to MSD. In addition, the predictive validity of these five EK FCE lifting tests on return to work is evaluated. In chapter 7, the main findings from the studies listed in this thesis are discussed, and conclusions are drawn regarding the clinimetric properties of the EK FCE tests. Furthermore, general aspects and recommendations for future research are proposed.
References

28. Morales K. UK government to appoint occupational health “tsar” to reduce work related illness. BMJ. 2005;331:986
34. Strong S. Functional capacity evaluation: the good, the bad and the ugly. OT Now 2002;Jan/Feb:5-9
44. Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. The complementary value of Functional Capacity Evaluation for physicians in assessing the physical work-ability of workers with musculoskeletal disorders. submitted
45. Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, Frings-Dresen MHW. The effect of Functional Capacity Evaluation assessment on the judgment of insurance physicians of the physical work-ability in the context of disability claims. submitted
46. 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;79:505-513