Int Arch Occup Environ Health (2004) 77: 527–537
DOI 10.1007/s00420-004-0549-7

REVIEW

Vincent Gouttebarge · Haije Wind
P. Paul F. M. Kuijer · Monique H. W. Frings-Dresen

Reliability and validity of Functional Capacity Evaluation methods: a systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen work system

Received: 21 November 2003 / Accepted: 25 June 2004 / Published online: 9 November 2004

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.

V. Gouttebarge (✉) · H. Wind · P. P. F. M. Kuijer
M. H. W. Frings-Dresen
Coronel Institute for Occupational and Environmental Health, AmCOGG: Amsterdam Centre for Research into Health and Health Care, Academic Medical Center/University of Amsterdam, PO Box 22700, 1100 DE, Amsterdam, The Netherlands
E-mail: v.gouttebarge@amc.uva.nl
Tel.: +31-20-5665337
Fax: +31-20-6977161

Keywords Functional Capacity Evaluation · FCE · Reliability · Validity · Systematic review

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 (Ferrari and Russel 2003; Marras 2000; Palmer 2000; Reginster 2002).

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 (IEBS, Statistic Central Desk 2001). 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 (Statistic Central Desk). 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 (Statistic Central Desk).

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 (Strong 2002; Tuckwell et al. 2002; Vasudevan 1996). FCEs are commonly used for individuals who have work-related disorders, particularly musculoskeletal disorders (Lechner 1998; Vasudevan 1996). 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 (Harten 1998; Innes and Straker 2002).

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), (Blankenship 1994) Ergos work simulator (EWS) Reference EWS FCE, Ergo-Kit (EK) (EKFCE 2002) and Isernhagen Work System (IWS) (IWSFCE).

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 (King et al. 1998). The providers of these FCEs pretend that these assessments use procedures that are reliable and valid (Lindeboom et al. 2003). 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 (Gardener and McKenna 1999).

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?

Materials and 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).

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

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.

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. Using one of the following FCEs: Blankenship, Ergos, Ergo-Kit, or Isernhagen
3. Presenting data about the reliability and/or validity of these FCEs.

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 (Altman 1991; Deyo et al. 1991; Innes and Straker 1999a, b; Numally 1978; Weiner and Stewart 1984).

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 (Deyo et al. 1991; Innes and Straker 1999a, b; Numally 1978; Weiner and Stewart 1984), and (5) *statistics* to evaluate whether the statistics used are clearly described and properly used to test the hypothesis of the study (Altman 1991).

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

- High: 4 or 5 +, indicating a high methodological quality
- Moderate: 2 or 3 +, indicating a moderate methodological quality
- 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 (Altman et al. 1991; Deyo et al. 1991; Innes and Straker 1999a, b; Numally 1978; Weiner and Stewart 1984). *n* number of subjects, *G* gender, *A* age, *H* health status, *W* work status

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
Objective	
+	The objective of the study is clearly mentioned
−	The objective of the study is not clearly mentioned
Population	
+	The five items *n*, G, A, H and W appear in the article
+/−	3–4 of the five items appear in the article
−	1–2 of the five items appear in the article
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 ten measurements
−	Number of raters used is 2 within ten 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
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 (Carmines and Zeller 1979; Streiner and Norman 2003). Different types of reliability are known as intrarater reliability, test–retest reliability, interrater reliability or internal consistency (Innes and Straker 1999a). 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 scores made by raters, testers or examiners on the same phenomenon (Innes and Straker 1999a). 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 (Portney and Watkins 2000).

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 (Carmines and Zeller 1979; Innes and Straker 1999b; King et al. 1998; Schultz-Jonhson 2002). In this study, we

Table 3 Levels of reliability and validity

Level of reliability: intrarater reliability, interrater reliability and internal consistency (Altman et al. 1991; Innes and Straker 1999a; Numally 1978)	
Pearson product moment coefficient (r), Spearman correlation coefficient (p), Somer correlation coefficient (d)[a]	
High	$r/p/d > 0.80$
Moderate	$0.50 \leq r/p/d \leq 0.80$
low	$r/p/d < 0.50$
Intra-class correlation coefficient ICC	
High	ICC > 0.90
Moderate	0.75 ≤ ICC ≤ 0.90
Low	ICC
Kappa value k	
High	$k > 0.60$
Moderate	$0.41 \leq k \leq 0.60$
Low	$k \leq 0.40$
Cronbach's alpha α	
High	$\alpha > 0.80$
Moderate	$0.71 \leq \alpha \leq 0.80$
Low	$\alpha \leq 0.70$
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 (Altman et al. 1991; Innes and Straker 1999b)	
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%, $\kappa > 0.60$, $r/d > 0.75$)[a]
Moderate	Some similarity between the test and the criterion measure (percentage agreement ≥70%, $\kappa \geq 0.40$, $r/d \geq 0.50$)[a]
Low	Little or no similarity between the test and the criterion measure (percentage agreement < 70%, $\kappa < 0.40$, $r/d < 0.50$)[a]
Construct validity: convergent and divergent validity	
High	Good ability to differentiate between groups or interventions, or good convergence/divergence between similar tests ($r \geq 0.60$)
Moderate	Moderate ability to differentiate between groups or interventions, or moderate convergence/divergence between similar tests ($r \geq 0.30$)
Low	Poor ability to differentiate between groups or interventions, or low convergence/divergence between similar tests ($r < 0.30$)

[a]Somer correlation coefficient (d) was ranged by the authors as the Pearson product moment coefficient (r) and Spearman correlation coefficient (p)

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 (Innes and Straker 1996b).

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) (Altman 1991; Innes and Straker 1999a, b; Numally 1978). 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, κ, 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.

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* (Boadella et al. 2003; Brouwer et al. 2003; Dusik et al. 1993; Gross and Battié 2001, 2003, 2004; IJmker et al. 2003; Isernhagen et al. 1999; Matheson et al. 2002; Reneman et al. 2001, 2002a, b , c; Rustenburg et al. 2004), and six reviews (Gibson and Strong 1997; Jones and Kumar 2003; King et al. 1998; Lechner 2002; Schultz-Johnson 2002; Tramposh 1992). 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%).

Methodological quality appraisal

During the methodological quality appraisal, two of the 14 papers were excluded. Boadella et al. (2003) did not

Table 4 Results of the methodological quality appraisal and the overall methodological quality

Authors	FCE method	Objective	Population	Procedure	Statistics	Methodological quality
Brouwer et al. (2003)	+	+	+	+	+	High
Dusik et al. (1993)	+	+	+/−	+/−	+	Moderate
Gross and Battié (2001)	+	+	+	+/−	+	High
Gross and Battié (2003)	+	+	+	+	+/−	High
Gross and Battié (2004)	+	+	+	+	+	High
IJmker et al. (2003)	+	+	+	+	+/−	High
Isernhagen et al. (1999)	+	+	+/−	+	+	High
Matheson et al. (2002)	+	+	+/−	+	+	High
Reneman et al. (2002a)	+	+	+/−	+/−	+/−	Moderate
Reneman et al. (2002b)	+	+	+	−	+	Moderate
Reneman et al. (2002c)	+	+	+	+	+/−	High
Rustenburg et al. (2004)	+	+	+	+/−	+/−	Moderate

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. (2001) 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

Table 5 Overview of the included studies on the reliability of the four FCE methods. *ICC* intra-class correlation coefficient, % percentage of agreement. *n* number of subjects/*G* gender/*A* age/*H* health status/*W* work status

FCE method (subtests)	Objective: type(s) of reliability	Population	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	$0.75 \leq ICC \leq 0.87$	Brouwer et al. (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≥0.78	Gross and Battié (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 $\kappa = 0.68$ (2) Judging lifting as light or heavy $\kappa = 0.81$	Isernhagen et al. (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 et al. (2002a)
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 = 0.87 (2) ICC = 0.87 (3) ICC = 0.77	Reneman et al. (2002b)

Table 6 Overview of the included studies on the validity of the four FCE methods. *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. *k* kappa value; *r* Pearson Correlation Coefficient; *p* Spearman's rank correlation; *d* Somer's coefficient

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) $\kappa = 0.629$ for overall $0.45 \leq r \leq 0.87$ for strength variables (2) $\kappa = 0.407$ (3) $\kappa \leq 0.45$	Dusik et al. 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) $r = -0.51$ (2) $r = -0.45$	Gross and Battié 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 and Battié 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	$r = 0.72$	IJmker et al. 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 $P < 0.005$ for return to work	Matheson et al. 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) $p = -0.17$ & $-.20/d = 0.03$ (2) $-0.08 \leq d \leq 0.23$ (3) $-0.52 \leq p \leq -0.27$ $-0.15 \leq d \leq 0.05$	Reneman et al. [43] 2002
Ergos WS	Concurrent validity	*n*: 25 subjects	Time interval of 7 days between assessments on EWS and EK (order FCE counter balanced)	$0.49 \leq p \leq 0.66$	Rustenburg et al. 2004
4 static and 6 dynamic lifting tests Ergo-Kit 4 lifting tests		G: 25 males A: 34.8 years H: healthy W: fire fighters			

ranked as high (Brouwer et al. 1993; Gross and Battié 2001, 2003, 2004; IJmker et al. 2003; Isernhagen et al. 1999; Matheson et al. 2002; Reneman et al. 2002c), and four as moderate (Dusik et al. 1993; Reneman et al. 2002a, b; Rustenburg et al. 2004).

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 (Dusik et al. 1993; Reneman et al. 2002a). For all of them, we did not find that high quality procedures were used to achieve their objectives: three were scored as moderate (Dusik et al. 1993; Reneman et al. 2002a; Rustenburg et al. 2004) and one as low (Reneman et al. 2002b). 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 (Dusik et al. 1993). Concerning the concurrent validity of the EWS and EK, the time interval between assessments on both FCEs was considered too long (Rustenburg et al. 2004). Concerning the intrarater reliability studies of the IWS, the time interval between test and retest was too short or too long (Reneman et al. 2002a, b).

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 (Brouwer et al. 2003; Gross and Battié 2001; Isernhagen et al. 1999), one on the concurrent validity of the IWS and EK (IJmker et al. 2003) and four on the predictive and concurrent validity of the IWS (Gross and Battié 2003, 2004; Matheson et al. 2002; Reneman et al. 2002c).

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.

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 EWS (Dusik et al. 1993; Rustenburg et al. 2004). Dusik et al. (1993) 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 ($\kappa = 0.63$) but poorly with the SHOP and VALPAR ($\kappa < 0.45$). 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. (2004) 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) (Rustenburg et al. 2004) and one on the concurrent validity of the EK and the IWS (IJmker et al. 2003). In this study, IJmker et al. (2003) 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 (Brouwer et al. 2003; Gross and Battié 2001; Reneman et al. 2002a, b), and three outcomes about the interrater reliability (Reneman et al. 2002a; Gross and Battié 2001; Isernhagen et al. 1999).

Four studies evaluated the intrarater reliability (test–retest) of the IWS. Brouwer et al. (2003) 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 (2001) 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≥0.78). Reneman et al. (2001a, b) 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 (Reneman et al. 2001a) and an intraclass correlation coefficient (Reneman et al. 2001b) 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 (Carmines and Zeller 1979; Matheson et al. 1996). 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 (2001) 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 (Fleiss 1986; Portney and Watkins 2000; Tinsley and Weiss 1975), and was rated, according to our scale, as high (all ICC≥0.95). This result is in line with the findings reported by Isernhagen et al. (1999). 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 ($\kappa = 0.81$). Reneman et al. (2002a) 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 (Gross and Battié 2003), two the predictive validity (Gross and Battié 2004; Matheson et al. 2002) and two the concurrent validity (IJmker et al. 2003; Reneman et al. 2002c).

IJmker et al. (2003) studied the concurrent validity of the IWS and the EK and the results are reported beforehand (see EK). Reneman et al. (2002c) 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 (2004) 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. (2002) 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 (2003) 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.

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 (Hart et al. 1993). 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 (Matheson et al. 1996, 2002; Portney and Watkins 2000). If an FCE measurement is not reliable, tests results are not consistent and it would be thus impossible to demonstrate its validity (King et al. 1998). Therefore, any study concerning the validity of one of the four FCEs should refer to or mention its reliability. Dusik et al. (1993), IJmker et al. (2003) and Rustenburg et al. (2004) 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 (Gross and Battié 2003, 2004; Matheson et al. 2002; Reneman et al. 2002c), 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 (Portney and Watkins 2000; Streiner and Norman 2003). In the studies of Dusik et al. (1993), IJmker et al. (2003) and Rustenburg et al. (2004), 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 (Carmines and Zeller 1979; Portney and Watkins 2000; Streiner and Norman 2003). In their studies, Dusik et al. (1993) and Rustenburg et al. (2004) 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 evalutions 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 judgement 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 counsellors 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.

References

Altman DG (1991) Practical statistics for medical research. Chapman and Hall, London

Blankenship KL (1994) The Blankenship system functional capacity evaluation: the procedure manual. The Blankenship Corporation, Macon

Boadella JM, Sluiter JK, Frings-Dresen MHW (2003) Reliability of upper extremity tests measured by the Ergos^{TM} work simulator: a pilot study. J Occup Rehabil 13:219–232

Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JMH, Göeken 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

Carmines EG, Zeller A (1979) Reliability and validity assessment. Sage Publications, Iowa

Statistic Central Desk. http://www.cbs.nl [Centraal Bureau voor Statistiek, in Dutch]:

Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures. Control Clin Trials 12:142S–158S

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

EKFCE (2002) Ergo-Kit functional capacity evaluation: User manual. Enschede, The Netherlands: Ergo Control, 2002. [Ergo-Kit Functionele Capaciteit Evaluatie. Handleiding, in Dutch]

EWS FCE: Ergos Work Simulator. Users Guide. Work Recovery System Inc., Tucson, Arizona

Ferrari R, Russel AS (2003) Regional musculoskeletal conditions. Best Pract Res Clin Rheumatol 17:57–70

Fleiss JL (1986) The design and analysis of clinical experiments. Wiley, New York

Gardener L, McKenna K (1999) Reliability of occupational therapists in determining safe, maximal lifting capacity. Aust Occup Ther J 46:110–119

Gibson L, Strong J (1997) A review of functional capacity evaluation practice. Work 9:3–11

Gross DP, Battié MC (2001) Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther 82:364–371

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

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

Hart DL, Isernhagen SJ, Matheson LN (1993) Guidelines for functional capacity evaluation of people with medical conditions. J Orthop Sport Phys 18:682–686

Harten JA (1998) Functional capacity evaluation. Occup Med State Art Rev 13:209–212

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

Innes E, Straker L (1999a) Reliability of work-related assessments. Work 13:107–124

Innes E, Straker L (1999b) Validity of work-related assessments. Work 13:125–152

Innes E, Straker L (2002) Workplace assessments and functional capacity evaluations: current practices of therapists in Australia. Work 18:51–66

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

IWSFCE: Isernhagen Work System Functional Capacity Evaluation. Manual. Duluth, Minnesota, USA

Jones T, Kumar S (2003) Functional capacity evaluation of manual materials handlers: a review. Disabil Rehabil 25:179–191

King PM, Tuckwell N, Barrett TE (1998) A critical review of functional capacity evaluations. Phys Ther 78:852–866

Lechner DE (1998) Functional capacity evaluation. In: King PM (ed) Sourcebook of occupational rehabilitation. Plenum, New York, pp 209–227

Lechner DE (2002) The role of functional capacity evaluation in management of foot and ankle dysfunction. Foot Ankle Clin N Am 7:449–476

Lindeboom D, Bachoe S, Karsemeijer E, Faber L (2003) De plaats van FCE op gebied van onderzoek naar arbeidsgebonden problematiek, revalidatie en integratie. Rapport Arbeidsreïntegratie, Hulpmiddelen en Ergonomie [In Dutch]

Marras WS (2000) Occupational low back disorder causation and control. Ergonomics 43:880–902

Matheson LN, Mooney V, Grant JE, Legget S, Kenny K (1996) Standarized evaluation of work capacity. J Back Musculoskelet 6:249–264

Matheson LN, Isernhagen SJ, Hart DL (2002) Relationships among lifting ability, grip force, and return to work. Phys Ther 82:249–256

Mooney V (2002) Functional capacity evaluation. Orthopedics 25:1094–1099

Numally JC (1978) Psychometric theory, 2nd edn. McGraw-Hill, New York

Palmer KT (2003) Pain in forearm, wrist and hand. Best Pract Res Clin Rheumatol 17:113–135

Portney LG, Watkins MP (2000) Foundations of clinical research: applications to practice. Appleton and Lange, Norwalk

Reginster JY (2002) The prevalence and burden of arthritis. Rheumatology 41(Suppl 1):3–6

Reneman MF, Joling CI, Soer EL, Göeken LNH (2001) Functional capacity evaluation: ecological validity of three static endurance tests. Work 16:227–234

Reneman MF, Jaegers SMHJ, Westmaas M, Göeken LNH (2002a) The reliability of determining effort level of lifting and carrying in a functional capacity evaluation. Work 18:23–27

Reneman MF, Dijkstra PU, Westmaas M, Göeken LNH (2002b) Test–Retest reliability of lifting and carrying in a 2-day functional capacity evaluation. J Occup Rehabil 12:269–275

Reneman MF, Jorritsma W, Schellekens JMH, Göeken LNH (2002c) Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic non-specific low back pain. J Occup Rehabil 12:119–129

Rustenburg G, Kuijer PPFM, Frings-Dresen MHW (2004) The concurrent validity of the ERGOS^{TM} work simulator and the Ergo-Kit$^{®}$ with respect to maximum lifting capacity. J Occup Rehabil 14:107–118

Schultz-Jonhson K (2002) Functional capacity evaluation following flexor tendon injury. Hand Surg 7:109–137

Streiner DL, Norman GR (2003) Health measurement scales. Oxford University Press, New York

Strong S (2002) Functional capacity evaluation: the good, the bad and the ugly. Occup Ther Now 5–9

Tinsley HEA, Weiss DJ (1975) Interrater reliability and agreement of subjective judgements. J Couns Psychol 22:358–376

Tramposh AK (1992) The functional capacity evaluation: measuring maximal work abilities. Occup Med State Art 7:113–124

Tuckwell NL, Straker L, Barrett TE (2002) Test–retest reliability on nine tasks of the physical work performance evaluation. Work 19:243–253

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]

Vasudevan SV (1996) Role of functional capacity assessment in disability evaluation. J Back Musculoskelet 6:237–248

Weiner EA, Stewart BJ (1984) Assessing individuals. Little Brown, Boston