Journal of Occupational Rehabilitation, Vol. 14, No. 4, December 2004 (© 2004)

Test–Retest Reliability of the Isernhagen Work Systems Functional Capacity Evaluation in Healthy Adults

M.F. Reneman,[1,2,5] **S. Brouwer,**[1,3] **A. Meinema,**[1] **P.U. Dijkstra,**[1,3,4] **J.H.B. Geertzen,**[1,3] **and J.W. Groothoff**[3]

Aim of this study was to investigate test–retest reliability of the Isernhagen Work System Functional Capacity Evaluation (IWS FCE) in healthy subjects. The IWS FCE consists of 28 tests that reflect work-related activities such as lifting, carrying, bending, etc. A convenience sample of 26 healthy subjects participated in the study. The subjects' mean age was 34.9 years. Two FCE sessions were held within a 2–3 week interval. Descriptives per session, Intra Class Correlations (ICC), limits of agreement, Cohen's Kappa, and percentage of agreement were calculated where appropriate. An ICC of ≥0.75, a Kappa value ≥0.60, and a percentage of absolute agreement of ≥80% were considered acceptable reliability. Acceptable reliability was demonstrated for seven out of nine tests (78%) of the material handling group and the shuttle walk test based on ICC analyses only. Sixteen out of 17 criterion and ceiling tests (94%) showed acceptable reliability based on Kappa values and percentage of agreement. Of these 17 tests, 8 were eligible for further analysis, and of those 8 tests the reliability of one test was acceptable based on ICC analyses (13%). In conclusion, the test–retest reliability of the material-handling group is acceptable. Crude analyses of the ceiling and criterion tests reveal acceptable test–retest reliability of most, but not all, tests.

KEY WORDS: reliability; functional capacity evaluation; disability assessment; occupational rehabilitation.

INTRODUCTION

Functional Capacity Evaluations (FCEs) are test batteries aimed at measuring the ability of a person to perform work-related activities. FCEs are based upon the job factors of the Dictionary of Occupational Titles, a publication of the United States Department of Labor (1–3). This dictionary describes the physical activities (job factors) that a job requires in a systematic way, by means of physical demands analysis. Many FCEs are available on

[1]Center for Rehabilitation, University Hospital Groningen, The Netherlands.
[2]Center for Occupational Health, University Hospital Groningen, The Netherlands.
[3]Northern Center for Health Care Research, University of Groningen, The Netherlands.
[4]Department of Oral and Maxillofacial Surgery, University Hospital Groningen, The Netherlands.
[5]Correspondence should be directed to M.F. Reneman, Center for Rehabilitation, University Hospital Groningen, PO Box 30.002, 9750 RA Haren, The Netherlands; e-mail: m.reneman@beatrixoord.nl.

1053-0487/04/1200-0295/0 © 2004 Springer Science+Business Media, Inc.

the market, one of which is the Isernhagen Work System (IWS) FCE. The IWS FCE consists of 28 tests that measure work-related activities (Table I).

Different aspects of the IWS FCE already have been tested for their reliability. In a test–retest design "lifting" and "carrying" have been found to possess a good reliability, with intraclass correlations ranging from 0.77 to 0.94 (4,5). Static pushing and static pulling also appeared to possess good test–retest reliability (6), as does the measurement of maximum holding times (7). Test–retest reliability of almost all tests of the IWS FCE was recently investigated in a sample of patients suffering from chronic nonspecific low back pain (8). The reliability of many tests of the IWS FCE was deemed acceptable, indicated by different statistical indices. This means that at group level, the results of the first testing session did not differ significantly from the second session. One of the important findings was the large variance between the test sessions, as indicated by large limits of agreement. This means that at the level of the individual patient, the performances could differ substantially between sessions. The source of variation in performances may be attributed to the patient behavior, properties of the testing protocols, or to variation related to the evaluator.

This study was conducted to investigate test–retest reliability of almost all tests of the IWS FCE on healthy subjects, and to compare the variation with the variation of patients with CLBP (8). In comparison to the study by Brouwer *et al*. (8) most of the methodology of this study and analysis of the results were held constant, however, healthy subjects participated in this study instead of patients with CLBP in the Brouwer study.

METHODS

Subjects

Subjects were recruited on the basis of convenience. All declared to be healthy, i.e. to have no medical condition that would restrict them from performing maximally. The original sample consisted of 28 subjects. Two subjects were unable to perform the second session because of acute low back pain unrelated to the first testing session. The study sample thus consisted of 26 subjects: 14 males and 12 females. Their mean age was 34.9 years (SD 12.7 years), mean weight was 83.5 kg (SD 15.5), and mean length was 181 cm (SD 8.5). One of the subjects experienced an episode of acute low back pain after the first session, and performed only half of the items of the second session.

Procedures

Two testing sessions were held within a 2–3 week interval. Time of day was kept constant. The subjects were introduced to the general FCE procedures per IWS FCE protocol (9) and then signed informed consent. Prior to each test, the subjects were briefly instructed verbally on the required performance. The evaluator then demonstrated each test. In this way, a total of 28 tests were performed (Table I). The subjects were asked to perform to their maximum abilities. Testing could be terminated for four reasons. 1) It was explained that they were allowed to stop the procedures at any point if they wished to do so, for example because of insecurity or pain. 2) The subjects wore a heart rate monitor throughout the test

Table I. Description of the Activities of the Isernhagen Work Systems (IWS) FCE

FCE activity	Description	Scoring
Lifting low	5 lifts from table to floor v.v.; 4–5 weight increments; <90 s	Max amount kg lifted
Overhead lift	5 lifts from table to crown height v.v.; 4–5 weight increments; <90 s	Max amount kg lifted
Short carry two handed	5 carries 1.5 m; waist height; 4–5 weight increments; <90 s	Max amount kg carried
Long carry two handed	1 carry 20 m; waist height; 4–5 weight increments; <90 s	Max amount kg carried
Long carry right handed	1 carry 20 m; waist height; 4–5 weight increments; <90 s	Max amount kg carried
Long carry left handed	1 carry 20 m; waist height; 4–5 weight increments; <90 s	Max amount kg carried
Pushing static	Static full body push; 3 repetitions	average kgF
Pulling static	Static full body pull; 3 repetitions	average kgF
Pushing dynamic	Pushing a weighted cart over 10 m including 2 turns	Safely possible yes/no
Pulling dynamic	Pulling a weighted cart over 10 m including 2 turns	Safely possible yes/no
Overhead work test*	Standing with hands at crown height; manipulating nut/bolts, max. 15 min	Time position is held (s)
Forward bend test standing*	Standing with 30–60° trunk flexion; manipulating nut/bolts, max. 15 min	Time position is held (s)
Forward bend test sitting	Sitting with 30–60° trunk flexion; manipulating nut/bolts, max. 5 min	Time position is held (s)
Kneeling	Maintaining kneeling posture; knees 90° flexion, hips straight, max. 5 min	Time position is held (s)
Crawling	Ambulate 3 m on hands and knees, then replace small object from floor to table height while in crawling position; 10 reps	Able yes/no
Crouching	Maintaining position with knees and hips fully flexed, max 1 min	Time position is held (s)
Dynamic bending*	Repetitive bending at hips and back; remove small object from floor to crown height 20 reps	Time needed to perform 20 reps (sec)
Dynamic squatting	Repetitive squatting with full flexion at knees and hips; remove small object from floor to crown height 20 reps	Time needed to perform 20 reps (sec)
Rep. rotation standing right*	Remove object horizontally at table height from left to right with left hand/arm; distance wing span; 30 reps; standing	Time needed to perform 30 reps (sec)
Rep. rotation standing left*	Remove object horizontally at table height from right to left with right hand/arm; distance: wing span; 30 reps; standing	Time needed to perform 30 reps (sec)
Rep. rotation sitting right*	Remove object horizontally at table height from left to right with left hand/arm; distance wing span; 30 reps; sitting	Time needed to perform 30 reps (sec)
Rep. rotation sitting left*	Remove object horizontally at table height from right to left with right hand/arm; distance: wing span; 30 reps; sitting	Time needed to perform 30 reps (sec)
Walking*	Shuttle walk test; increase speed per minute	Highest level completed
Stairclimbing	Ascend and descent 100 steps; no handrail	Able yes/no
Ladder climbing*	Ascend and descent stepladder with 5 steps with use of hands	Able yes/no
Balance	Walking over a 10 × 300 cm balance board; forward, backward, heel to toe, sideways (6 ways; total mistakes)	Able with less than 6 mistakes yes/no

Note. Rep: repetitive; max: maximal; s: seconds; kg: kilograms; kgF: kilograms force; v.v.: vice versa; m: meters; Tests marked (*) are modified from the standard IWS FCE protocol.

procedures. A test was terminated when the subject's heart rate met or exceeded 85% of his or her age-related maximum. 3) The evaluator terminated testing if it became unsafe, defined as a situation in which the subject was not in full control of him- or herself and/or the load. 4) For some tests a predetermined time limit was the reason for the subject to stop (i.e. crouching, max. 60 s). The evaluator recorded the results directly after each test. One internally trained evaluator evaluated all subjects. Each session lasted approximately 2 h.

A modified IWS FCE was used in this study. Instead of a 2-day protocol suggested in the original IWS protocol, all tests were performed on a single day (5). The tests pushing and pulling dynamic, crawling, walking, stair climbing, and ladder climbing were slightly modified (described and marked * in Table I). Minor modifications were also made to the following tests. The overhead work test and the forward bend test standing: patients were instructed to hold these postures as long as possible (7). The ceiling of these tests was set at 15 instead of 5 min because otherwise too many subjects reached this ceiling and would not perform to their maximum capacities. Sitting and standing tolerances were excluded from this study because of the duration of these tests (each test lasts 30 min) and because of the fact that most CLBP patients are consistently able to tolerate these tests (8). As was the case in the Brouwer study, the grip strength and hand coordination tests were excluded from the study protocol.

Data Analysis

In the IWS FCE three types of tests can be distinguished: those with a criterion, those with a ceiling, and those without criterion or ceiling (i.e., material handling tests and shuttle walk test). For the material handling tests and shuttle walk test means, standard deviations, 95% confidence intervals, intraclass correlations (ICC, model one way random), and limits of agreement were calculated (10).

A criterion indicates that a test is fulfilled when a criterion is met. For instance the test "pushing dynamic" has a criterion that a subject is able or unable to safely push a weighted cart over a distance of 20 m. The repetitive rotation tests are criterion tests as well; the time needed to perform 30 rotations to the left and the right side. For all criterion tests the number of subjects who met the criterion for each test session was calculated. On the basis of these dichotomous results Cohen's Kappa's and percentages of absolute agreement of subjects with identical test behavior over two test sessions were calculated. Cohen's Kappa's could not be calculated when the filling of the 2×2 tables was incomplete. If a subject reached the criterion in sessions 1 and 2, the times needed to reach the criterion in the sessions were used for further analyses: means, standard deviations, 95% confidence intervals, ICCs, and limits of agreement were calculated. Data of subjects not meeting the criterion in session 1 or 2 were excluded from further analyses.

A ceiling indicates that the test is terminated because the subject has met what is defined to be the maximal time of performance. For instance, working static overhead has a ceiling of 15 min. This test was terminated when the subject reached 15 min. In that case maximal capacity was not reached. For all tests with a ceiling effect, the number of subjects who reached the ceiling for each test session was calculated. On the basis of these dichotomous results Cohen's Kappa's and percentages of absolute agreement of subjects with identical test behavior over the two sessions were calculated. Cohen's Kappa's could

not be calculated when the filling of the 2 × 2 tables was incomplete. If a subject reached the ceiling in session 1 or 2, the data of that subject were excluded for further analyses because this subject's maximal performance could not be analyzed. Of the subjects who did not reach the ceiling in both sessions means, standard deviations, 95% confidence intervals, ICCs, and limits of agreement were calculated.

Criteria for interpretation of the indices for reliability were equal to those described in the study of Brouwer *et al*. (8). An ICC of 0.75 or more was considered acceptable reliability. A Kappa value of more than 0.60 was considered an acceptable reliability. Arbitrarily, a percentage of absolute agreement of 80% or more was also considered an acceptable reliability. All analyses were performed in SPSS.

RESULTS

Material-Handling Group and Shuttle Walk Test

The results of the test–retest reliability of the material-handling tests yielded ICC values ranging from 0.68 to 0.98 and limits of agreement ranging from 4.8 to 21.5 kg (Table II). Limits of agreement could not always be calculated because there was a systematic difference between the first and the second session (10). Mean performances were generally better on the second testing session. Seven out of eight tests reached ICC values higher than 0.75. The ICC value of the shuttle walk test was 0.64 with limits of agreement of ±199.6 m.

Table II. Results of Paired *t* Test, Limits of Agreement and ICC's of the Material Handling Tests of the Modified Isernhagen Work System FCE and the Shuttle Walk Test

Activity (*n* paired observations)	Mean 1	SD 1	Mean 2	SD 2	Mean difference	SD	95% CI of difference	Limits of agreement	ICC	95% CI of ICC
Lifting low in kg (25)	45.8	17.6	50.2	20.5	−4.4	4.5	−6.3 to −2.5	—	0.95	0.89–0.98
Lifting high in kg (25)	19.6	6.2	20.5	5.7	−0.9	2.7	−2.0 to 0.3	±5.6	0.89	0.77–0.95
Carry short in kg (25)	47.3	18.7	53.4	20.7	−6.0	6.8	−8.9 to −3.2	—	0.90	0.78–0.95
Carry long in kg (25)	46.4	15.5	50.9	15.1	−4.5	7.6	−7.6 to −1.3	—	0.84	0.68–0.93
Carry right in kg (25)	37.2	11.9	38.5	11.4	−1.2	2.3	−2.2 to −0.3	—	0.98	0.95–0.99
Carry left in kg (25)	35.1	9.8	36.2	10.6	−1.1	1.1	−3.4 to 1.2	±4.8	0.86	0.70–0.93
Pushing static in kg (26)	40.9	12.8	47.3	16.8	−6.5	10.5	−10.7 to −2.2	—	0.68	0.41–0.84
Pulling static in kg (26)	56.0	17.7	58.9	19.4	−2.9	8.5	−6.3 to 0.6	±21.5	0.89	0.77–0.95
Shuttle walk test in meters (25)	517.2	112.2	555.6	123.3	−38.4	96.7	−77.9 to 1.1	±199.6	0.64	0.34–0.82

Note. Mean 1: Group mean in the first session; Mean 2: Group mean in the second session; ICC: Intraclass correlation (one way random model); 95% CI: 95% Confidence interval; — Limits of agreement could not be calculated because there is a systematic difference between the first and the second session.

Table III. Criteria and Ceiling for Test Termination, Kappa's and Percentage of Similar Test Behavior, for Different Tests of the Modified Isernhagen Work System FCE

Tests (*n* paired observations)	Statistical level	Criterion*/Ceiling**	*n* Subjects reaching criterion or ceiling in session 1	*n* Subjects reaching criterion or ceiling in session 2	Kappa (κ)	Similar test behavior
Pushing dynamic (19)	Dichotomous	20 m*	19	19	#	100% (19/19)
Pulling dynamic (19)	Dichotomous	20 m*	19	19	#	100% (19/19)
Overhead work test (26)	Continuous	15 min**	2	3	0.78	96% (25/26)
Forward bend test standing (24)	Continuous	15 min**	6	6	1.00	100% (24/24)
Forward bend test sitting (24)	Continuous	5 min**	14	15	0.57	79% (19/24)
Kneeling (24)	Continuous	5 min**	23	22	0.65	96% (23/24)
Crawling (26)	Dichotomous	10 repetitions*	26	26	#	100% (26/26)
Crouching (20)	Dichotomous	60 s*	17	18	0.77	95% (19/20)
Dynamic bending (26)	Continuous	20 repetitions*	26	26	#	100% (26/26)
Dynamic squatting (25)	Continuous	20 repetitions*	25	25	#	100% (25/25)
Rotation standing right (25)	Continuous	30 repetitions*	25	25	#	100% (25/25)
Rotation standing left (25)	Continuous	30 repetitions*	25	25	#	100% (25/25)
Rotation sitting right (25)	Continuous	30 repetitions*	25	25	#	100% (25/25)
Rotation sitting left (25)	Continuous	30 repetitions*	25	25	#	100% (25/25)
Stair climbing (26)	Dichotomous	20 × 5 steps up/down*	14	16	0.69	85% (22/26)
Ladder climbing (25)	Dichotomous	5 times up/down*	25	25	#	100% (25/25)
Balance (total of 6 tests) (24)	Dichotomous	less than 6 failures*	25	24	#	96% (24/25)

*Criterion indicates that a test is fulfilled if the criterion is met. For instance the test pushing dynamic has as criterion that a subject is able or not able to push a weighted cart over a distance of 20 m safely.

**Ceiling indicates that the test is terminated because the patient has met what is defined to be the maximal time of performance. For instance, working static overhead has a ceiling effect at 15 min. The test is terminated when the subject reaches 15 min. However, in that case he/she has not performed to his/her maximal ability.

#Kappa values can not be calculated because of lack of filling of the cells in the 2 × 2 table.

Criterion and Ceiling Tests

The results of the reliability of tests with a ceiling or a criterion are presented in Tables III and IV. As shown in Table III, similar test behavior was seen in all 17 tests, as indicated by percentages of agreement exceeding 80%. Kappa values of 0.60 or higher were found for five out of six tests in which a Kappa could be calculated. Kappa could not be calculated in 11 tests because of incomplete filling of the cells in the 2 × 2 tables. The results of the additional analyses of the tests with a ceiling or a criterion are presented in Table IV. During the second testing session, the subjects performed worse on the ceiling tests, and performed better on most criterion tests. No further analyses were performed

Table IV. Results of Paired *t* Test, Limits of Agreement and ICC's for Tests (With a Criterion* or a Ceiling** Effect) of the Modified Isernhagen Work System FCE

Activity in seconds (*n* paired observations)	Mean 1	SD 1	Mean 2	SD 2	Mean difference	SD	95% CI	Limits of agreement	ICC	95% CI
Working static overhead** (23)	403.8	181.9	348.8	120.8	55.0	135.9	−3.75–113.8	±281.8	0.58	0.23–0.79
Forward bend test standing** (18)	405.1	174.1	377.3	197.12	27.7	117.8	−30.9–86.3	±248.6	0.93	0.85–0.97
Dynamic bending* (19)	50.5	7.0	47.5	5.5	3.1	6.1	0.6–5.6	—	0.45	0.09–0.71
Squatting* (25)	42.5	5.1	41.6	3.3	1.0	4.1	−0.7–2.7	±8.5	0.54	0.20–0.77
Rotation standing right* (25)	70.4	8.7	70.9	9.5	−0.4	7.7	−3.6–2.7	±15.8	0.66	0.37–0.83
Rotation standing left* (25)	68.9	7.7	69.7	7.6	−0.8	5.6	−3.1–1.5	±11.6	0.73	0.49–0.87
Rotation sitting right* (25)	77.3	10.8	75.8	11.0	1.6	10.5	−2.8–5.9	±21.8	0.54	0.20–0.77
Rotation sitting left* (25)	76.2	9.8	74.0	8.3	2.3	6.6	−0.4–5.0	±13.5	0.72	0.47–0.87

Note. Mean 1: Group mean in the first session; Mean 2: Group mean in the second session; ICC: Intraclass correlation (one way random model); 95% CI: 95% Confidence interval; —: Limits of agreement can not be calculated because there is a systematic difference between the first and the second session.

*Only if a subject reached the criterion in sessions 1 and 2, the times needed to reach the criterion in the sessions were used for further analyses, other subjects were excluded from the analysis.

**If a subject reached the ceiling in session 1 or 2, he/she was excluded for further analyses because the maximal performance cannot be analyzed.

Table V. Summary and Interpretation of Results. "High" or "Low" Means the Reliability Coefficient is Higher or Lower than the Criteria for Interpretation (Kappa 0.60, 80% agreement, ICC 0.75)

FCE activity	Kappa	% Agreement	Reliability	ICC	Reliability
Lifting low	N/A	N/A	N/A	High	Acceptable
Overhead lift	N/A	N/A	N/A	High	Acceptable
Short carry two handed	N/A	N/A	N/A	High	Acceptable
Long carry two handed	N/A	N/A	N/A	High	Acceptable
Long carry right handed	N/A	N/A	N/A	High	Acceptable
Long carry left handed	N/A	N/A	N/A	High	Acceptable
Pushing static	N/A	N/A	N/A	Low	Not acceptable
Pulling static	N/A	N/A	N/A	High	Acceptable
Pushing dynamic	NC	High	Acceptable	N/A	N/A
Pulling dynamic	NC	High	Acceptable	N/A	N/A
Overhead work test*	High	High	Acceptable	Low	Not acceptable
Forward bend test standing*	High	High	Acceptable	High	Acceptable
Forward bend test sitting	Low	Low	Not acceptable	N/A	N/A
Kneeling	High	High	Acceptable	N/A	N/A
Crawling	NC	High	Acceptable	N/A	N/A
Crouching	High	High	Acceptable	N/A	N/A
Dynamic bending*	NC	High	Acceptable	Low	Not acceptable
Dynamic squatting	NC	High	Acceptable	Low	Not acceptable
Rep. rotation standing right*	NC	High	Acceptable	Low	Not acceptable
Rep. rotation standing left*	NC	High	Acceptable	Low	Not acceptable
Rep. rotation sitting right*	NC	High	Acceptable	Low	Not acceptable
Rep. rotation sitting left*	NC	High	Acceptable	Low	Not acceptable
Walking*	N/A	N/A	N/A	Low	Not acceptable
Stairclimbing	High	High	Acceptable	N/A	N/A
Ladder climbing*	NC	High	Acceptable	N/A	N/A
Balance	NC	High	Acceptable	N/A	N/A

Note. ICC: Intraclass Coefficient; N/A: Not applicable; NC: Not calculated
*Tests modified from the standard IWS FCE protocol.

on the following tests because the data of six or less subjects were eligible for analyses: pushing and pulling dynamic, forward bend test sitting, kneeling, crawling, crouching, stair climbing, ladder climbing, and balance.

All 28 tests of the IWC FCE were divided into tests with and tests without an acceptable reliability on the basis of the percentage of similar test behavior, the Kappa values, and ICCs. The results are presented in Table V. Summarizing the results of this study, seven out of nine tests (78%) of the material handling group and the shuttle walk test showed acceptable levels of reliability on the basis of ICC analyses only. Sixteen out of 17 criterion and ceiling tests (94%) showed acceptable reliability on the basis of Kappa values and percentage of agreement. Of these 17 tests, eight were eligible for further analysis, and of those eight tests the reliability of one was acceptable on the basis of ICC analyses (13%).

DISCUSSION

Material-Handling Group and Shuttle Walk Test

The ICC values of seven out of eight material handling tests were well over 0.75, thus these tests are reliable. The limits of agreement relative to the mean performances vary from 13 to 38% (limits of agreement/mean). The ICC values of the static pushing tests and

the shuttle walk test of the healthy subjects were below the criterion of 0.75, indicating these tests to be not reliable. It is unlikely that the substandard ICC value in this study can be attributed to a difference between or within the healthy subjects' performances. With regards to the static pushing test, the difference may be explained by a difference in evaluator. This particular test requires a specific subject performance to avoid peak forces. Our evaluator was internally trained and this particular aspect might have been undertrained. Test–retest reliability of this test was studied once before on 62 patients with chronic pain patterns by certified evaluators (6). The reliability in that study was ICC of 0.96 for pushing and 0.95 for pulling, which underscore that evaluator training may be of importance in this test. The reliability of the shuttle walk test is much lower in healthy subjects (ICC = 0.64) compared to CLBP patients (ICC = 0.84), patients with cardiac ($r = 0.99$, (11)) and pulmonary conditions ($r = 0.94$ to $r = 0.99$, (12)), and of healthy elderly people (13). The authors cannot explain this detonating finding satisfactorily.

Criterion and Ceiling Tests

Five out of six tests where a Kappa could be calculated were found reliable. A Kappa could not be calculated in 11 other tests because of lack of cell filling. In all of these 11 tests, the percentages of agreement was very high (96–100%), and are therefore considered reliable as well. On the basis of percentage agreement and Kappa values, all criterion and ceiling tests, with exception of the forward bend test sitting, are reliable for use in healthy subjects. Additional analyses of the data of those subjects who did not reach a ceiling or who met a criterion in session 1 or 2 were performed on eight tests. The forward bend test standing was the only test with acceptable reliability. The ICC values of all other seven tests that were analyzed were below 0.75, indicating that these tests are unreliable for use in healthy subjects.

The study design and analyses of this study were a replicate of the study described by Brouwer *et al*. (8). The discussion section of that study deals in part about theoretical considerations regarding the statistical analyses of the results. These considerations apply to this study as well. In summary, specifics in test design caused ceiling and floor effects, which in turn prevented a simple statistical analysis of the data. Additionally, a lack of cell filling prohibited calculation of Kappa statistics in 11 out of 17 applicable tests.

Comparing Variance of Healthy Subjects With CLBP Patients

A comparison of the results of this study with healthy subjects with the results of patients with CLBP reveals the following. The mean ICC value of the eight tests of the material-handling group is 0.87 in the healthy group and 0.81 in the CLBP group. The mean width of the standard deviations relative to the performances during the eight material handling tests (SD_{mean}/mean performances), was 34% in both sessions of the healthy subjects, and 38% in the first session and 42% in the second session of CLBP patients in the Brouwer study. Comparing percentages of similar test behavior to patients with CLBP revealed a mean percentage agreement of 96.9% in the healthy group and 91.8% in the CLBP group. The consistency between sessions of the healthy subjects' performances was better in eight tests, equal in six tests, and worse in three tests compared to CLBP patients.

Of the eight tests analyzed additionally, the ICC values of three tests were lower and of five tests were higher than in CLBP patients (8). Overall, higher ICC values, smaller limits of agreement, smaller standard deviations, and higher percentage agreement indicate less within subject variance of the healthy subjects. As hypothesized by Brouwer *et al*. (8), the large within subject variance of patients with CLBP can in part be attributed to the characteristics of the patients.

In conclusion, the test–retest reliability of the material-handling group is acceptable. Analyses of the ceiling and criterion tests reveal acceptable test–retest reliability of most, but not all, tests. Detailed analyses of these tests indicate levels of reliability that are unacceptable, indicating considerable within subject variances between occasions.

Work-related assessments should have demonstrated proof of safety, reliability, validity, practicality, and utility (14). No assessment can perform maximally on all five requirements; protocols are always a weighted balance between these requirements. For example, when considering the criterion of practicality, it can easily be explained why ceiling effects were created to reduce testing time. Introducing performance ceilings, however, does have its drawback: it is complicated to study the reliability of these tests to name just one. It is likely that contemplations such as these had taken place in the developmental stages of the IWS FCE, more than 20 years ago. Research has been performed and new data are now available about strengths and weaknesses of the standardized version of the IWS FCE. Design adjustments should be considered to improve the weaknesses of the current protocol.

ACKNOWLEDGEMENT

The authors thank the individuals who volunteered to participate in this study.

REFERENCES

1. King PM, Tuckwell N, Barrett TE. A critical review of Functional Capacity Evaluations. *Phys Ther* 1998; 78: 852–866.
2. Innes E, Straker L. Validity of work-related assessments. *Work* 1999; 13: 125–152.
3. Abdel-moty E, Fishbain DA, Khalil TM, Sadek S, Cutler R, Rosomoff RS, Rosomoff HL. Functional capacity and residual functional capacity and their utility in measuring work capacity. *Clin J Pain* 1993; 9: 2003–2013.
4. Gross DP, Battie MC. Reliability of safe maximum lifting determinations of a functional capacity evaluation. *Phys Ther* 2002; 82: 364–371.
5. Reneman MF, Dijkstra PU, Westmaas M, Göeken LNH. Test–retest reliability of lifting and carrying in a 2-day functional capacity evaluation. *J Occup Rehabil* 2002; 12: 269–275.
6. Hart DL. Test–retest reliability of the static push/pull tests for functional capacity evaluations. *Phys Ther* 1988; 68: 824.
7. Reneman MF, Bults MMWE, Engbers LH, Mulders KKG, Goeken LNH. Measuring maximum holding times and perception of static elevated work and forward bending in healthy young adults. *J Occup Rehabil* 2001; 11: 87–97.
8. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JMH, Göeken LNH. Test–retest reliability of the Isernhagen functional capacity evaluation in patients with chronic low back pain. *J Occup Rehabil* 2003; 13: 207–217.
9. Isernhagen Work Systems. *Functional capacity procedure manual*, 1st edn. Duluth, MN: Isernhagen Work Systems, 1997.
10. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 8: 307–310.
11. Morales FJ, Montemayor T, Marinez A. Shuttle versus six-minute walk test in the prediction of outcome in chronic heart failure. *Int J Cardiol* 2000; 76: 101–105.

12. Singh SJ, Morgan MDL, Scott S, Walters D, Hardman AE. Development of a shuttle walking test in chronic airflow limitation. *Thorax* 1992; 47: 1019–1024.
13. Lemmink KAPM. De Groninger fittest voor ouderen: Ontwikkeling van een meetinstrument. Ph.D. Thesis, Institute for Movement Sciences, University of Groningen, The Netherlands, 1996.
14. Hart DL, Isernhagen SJ, Matheson LN. Guidelines for functional capacity evaluations of people with medical conditions. *JOSPT* 1993; 18: 682–686.