Reliability of work-related assessments

Ev Innes[a,*] and Leon Straker[b]
[a]*School of Occupation and Leisure Sciences, Faculty of Health Sciences, University of Sydney. P.O. Box 170, Lidcombe, NSW 1825, Australia*
[b]*School of Physiotherapy, Faculty of Health Sciences, Curtin University of Technology, Selby St., Shenton Park, WA 6008, Australia*

Received 28 September 1998
Accepted 13 October 1998

Insufficient evidence of the reliability of work-related assessments is a major concern in this area of practice. Despite this concern there has been ongoing development of new assessments, while existing assessments have been revised, modified and updated and others are no longer used or available. *Objectives:* The purpose of this study was to determine the extent and quality of evidence for the reliability of work-related assessments. *Study Design:* This study examined available literature and sources in order to review the extent which reliability has been established for 28 work-related assessments. *Results:* The levels of evidence and reliability are presented for each assessment. This indicates that a number of commercially available work-related assessments have insufficient evidence of reliability. For the limited number of work-related assessments with an adequate level of evidence on which to judge their reliability, most demonstrate a moderate to good level. Few assessments, however, have demonstrated levels of reliability sufficient for clinical (and legal) purposes. *Conclusion:* With this study clinicians will be able to examine their options with regard to the reliability of the assessments they choose to use. Interpretation of changes in test results can be considered in the light of the evidence for the reliability of the instrument used.

Keywords: Reliability, work-related assessment, functional capacity evaluation

1. Introduction

Clinicians in the area of occupational rehabilitation have been encouraged to be informed regarding the extent to which reliability and validity have been established for work-related assessments, and how this information may impact on the interpretation of results obtained from these assessments [32]. For over a decade, numerous authors have identified insufficient evidence of reliability and validity of most work-related assessments as a major problem in this area of practice and research [1,23,27–29,32,34,37,38,66,69,70].

An attempt was made by Lechner et al. [38] to examine the existing evidence of reliability and validity for a number of common commercially available work-related assessments used in the United States (Blankenship FCE, BTE Work Simulator, Isernhagen FCE, Key FCA, Matheson, Physio-Tek, Polinsky FCA, Smith PCE, Sweat, Valpar CWS). The results from that review were of major concern. Of the assessments examined, it was reported that none had inter-rater reliability studies, and only one or two had intra-rater reliability, content, criterion-related or construct validity studies [38]. Unfortunately, the authors did not report the sources of these studies, and they also overlooked a number of published studies that had been conducted.

Since that time further development has occurred in the area, new assessments have been developed, existing assessments have been revised, modified and updated, and some assessments are no longer used or available. For this reason, the present study has examined current available literature and sources in order to review the extent to which reliability and validity have been established for a wide range of work-related assessments. Due to the extent of the information the results are presented in two articles – reliability in this article and validity in a subsequent paper.

1.1. Reliability

Reliability involves the extent to which a test or measure is consistent and free from error [51]. This consistency may be over time, between different raters, between equivalent forms of the same test, or between parts of the test [14]. In a clinical context this consistency is often in the form of actual agreement (i.e., obtaining the same results) and not just whether the results vary consistently and proportionally to each other.

*Corresponding author: Tel.: +61 2 9351 9209; Fax: +61 2 9351 9197; E-mail: E.Innes@cchs.usyd.edu.au.

Work 13 (1999) 107–124
ISSN 1051-9815 / $8.00

Confidence in the reliability of an instrument, testing procedure or individual making a measurement is crucial to clinicians when assessing clients, monitoring the efficacy of treatment and planning future intervention. If a measurement is reliable then changes noted in a client over time are likely to be due to real improvement or deterioration in the client's abilities, and not just due to measurement error.

1.2. Types of reliability

There are several types of reliability that are considered. The most common types associated with work-related assessments are test-retest or intra-rater reliability, and inter-rater reliability. Other types include equivalent, parallel or alternate forms, internal consistency and population-specific reliability.

Test-retest reliability determines the consistency of measures or scores from one testing occasion to another. It assumes that the characteristic being measured does not change over the time period. Test-retest reliability can be influenced by (1) testing effects, such as with practice or carryover effects of treatment, (2) rater bias, which can occur when subjective criteria are used to rate responses, and (3) test-retest intervals that must be far enough apart to avoid fatigue, memory or learning effects, and close enough to avoid genuine changes in performance [14,51]. Intra-rater reliability is considered to be a special case of test-retest reliability [51].

Rater reliability determines the consistency with which raters, testers or examiners make judgements, ratings or measurements of the same phenomenon [14]. In this situation the rater may (1) be the actual measurement instrument (e.g., manual muscle testing), (2) physically apply the tool and so become part of the instrument (e.g., measure range of motion with a goniometer), (3) observe performance and apply criteria to the subjective observation (e.g., evaluate lifting technique), or (4) read or interpret output from an instrument (e.g., grip strength from a dynamometer) [51]. *Intra-rater* reliability examines the stability of data recorded by one person across two or more testing occasions, while *inter-rater* reliability determines the variation between two or more raters who are assessing the same phenomenon.

Rater reliability is considered "especially important when measuring devices are new or unfamiliar or when subjective observations are used" [51, p. 60]. This is particularly the case with many work-related assessments that rely on clinical observations and judgements made by raters.

Equivalent, alternate or parallel form reliability involves comparing two or more different, but equivalent forms of the same test or measure. It is usually conducted when the assumptions regarding test-retest reliability cannot be met (i.e., subjects exhibit practice effects, rapid changes in the characteristic or are likely to recall responses) [14,51].

It has been noted that some forms of reliability, such as equivalent forms, split-half or internal consistency, are usually impractical to determine in functional assessments [52]. There is usually no equivalent form of the work-related assessment with which it can be compared, and Rosen [52, p. 47] notes that "there are so many factors underlying performance on work samples that the construction of truly equivalent forms may be only theoretically possible".

Internal consistency reflects the extent to which test items measure the same characteristic. Types of internal consistency are *split-half* and *test item* reliability. *Split-half reliability* is examined through comparing participants' results on equivalent halves of a test. It is used when a measure (e.g., a questionnaire) has more than one item. It may also be used when it is not possible to retest the same group of participants as would occur for test-retest reliability [14]. Some consider split-half to be superior to test-retest and equivalent form reliability because there is no time lag between testing and the mental, physical and environmental influences remain the same [51]. Split-half reliability may be determined on some work samples where multiple identical test items are produced, but this does not apply to many current work-related assessments.

Test item reliability "is an estimate of the consistency of one item with respect to other items on a measure" [14, p. 256]. Item analysis is used to examine how each test item relates to other items and the instrument as a whole [51]. Test item reliability may be more practical to determine than either alternate forms or split-half reliability in work-related assessments with multiple sub-tests. This would occur where a physical impairment, for example, is determined using multiple measurements, which then provides the basis for assessing the internal consistency of the testing procedures. Correlation between various test components can then be examined.

Generalisability of reliability results derived from one study to other therapists, environments, test conditions or subjects cannot be assumed unless these aspects are specifically addressed in the analysis. *Population-specific reliability* is an especially important issue, where many studies use non-injured or non-disabled

subjects to establish an assessment's reliability and then assume that this will apply to a patient population [47]. Factors such as pain, deformity, weakness and anxiety can alter an individual's response to measurement and the consistency of these measurements [51].

Intra-test reliability is also referred to by some researchers [42,43,46]. This is a form of response stability in which the consistency or stability of repeated measures over time are examined [51]. Intra-test reliability is determined from repeated measures taken within a short period of time (e.g., within 0.5 to 5 minutes of each other). The response stability is relevant when determining other aspects of reliability. If the measure or phenomenon is inherently unstable (i.e. varies widely each time a measure is taken) then it is not possible to determine whether errors in measurement are due to the instrument or the raters.

Instrument precision is an important area in which the instrument is examined for accuracy of measurement, rather than the reliability of clients being evaluated or clinicians using the instrument. The precision examined may be against a known and accepted standard (e.g., [12,19,20]) and so technically may be considered as a form of criterion-related validity [51], however, the accuracy of the instrument, and therefore the measurement obtained, is an important issue when determining all forms of reliability in clinical situations. The types of instrument precision examined in work-related assessments include inter- and intra-instrument reliability (e.g., [12,19,20]), which could be equated with inter-rater and test-retest reliability respectively. Response stability, or intra-test reliability, has also been examined (e.g., [60]), as has accuracy (e.g., [43,63]).

Test-retest, or intra-rater reliability is the most common form of reliability established for work-related assessments and Rosen [52] considers this to be usually the most sensible method. Hart [24], however, considers inter-rater reliability as the most important form of reliability, although he does note that good test-retest reliability is also critical.

The emphasis on the establishment of test-retest and inter-rater reliability in work-related assessments demonstrates the importance placed on ensuring that any change found on assessment is the result of change in the individual and not the result of measurement inconsistencies over time or between examiners. One area, however, which may benefit from further development is that of internal consistency. By examining the correlation between test items it may be possible to streamline assessment batteries to only include those items that assess necessary job requirements, rather than duplicating items which assess the same task or skill.

1.3. 'Good' Reliability

In order to determine if an assessment has adequate reliability for the purposes of clinical use, one must understand what measures are used to determine reliability, what are the strengths and limitations of these measures, and what the results that are produced actually mean from a clinical standpoint.

Reliability is usually reported as a coefficient. It is an estimate of the reliability based on variance, or the measure of differences among scores in a sample [51]. The range of scores used for estimating reliability must also be considered when interpreting the reliability coefficients. This is because reliability is based on "the proportion of the total observed variance that is attributable to error" [51, p. 508]. If the variance in scores is small (e.g., grip strength measurements with a range of 25 to 35 kg, $\bar{x} = 30.1, \mathrm{SD} = 3.414$), then an error (e.g., ± 2 kg) will be proportionally much larger than if the variance in scores was greater (e.g., grip strength measurements with a range of 15 to 45 kg, $\bar{x} = 30.1, \mathrm{SD} = 10.268$).

Many reliability coefficients (r) are based on measures of correlation (e.g., Pearson's Product-Moment (PPM) and Spearman Rank correlation coefficients). Correlation "reflects the degree of association between two sets of data, or the consistency of position within the two distributions" [51, p. 57]. Correlations reflect covariance rather than agreement between data sets. For many clinical applications, however, it is necessary to determine agreement between measures (i.e., two sets of measures are the same), not just whether they are proportional to each other [51]. Some consider that because correlations are unable to discriminate between the variance components due to error and those due to true differences, it is more accurate to use the *coefficient of determination* (r^2). This value reflects how much variance in one measurement is accounted for by the variance in the second measurement [51].

Correlations are also very sensitive to sample size. If a sample is large enough, almost any correlation coefficient will be found to be statistically significant. Therefore, the magnitude of the correlation coefficient should be considered when interpreting results, and not only whether the correlation is statistically significant [51].

Test-retest and rater reliability could use correlations (e.g., PPM or Spearman Rank) and t-tests to determine the consistency and average agreement respectively between data sets [51]. However, it is easier to interpret and more useful to have a single value to describe reliability. For that reason, it is preferable to use the

intra-class correlation coefficient (ICC) [51]. The ICC reflects both the degree of correspondence and agreement amongst the ratings. There are three models of the ICC (1, 2 and 3), and model 2 is considered the most appropriate to use when the aim is to demonstrate that the instrument can be used with confidence by all equally trained therapists [51].

When the unit of measurement in an assessment is categorical (e.g., poor, fair, good, excellent or safe/unsafe), reliability (usually inter-rater) is more appropriately determined as a measure of agreement. In its simplest form this would be *percentage agreement*, however, this value often gives an over-estimation of true reliability [51]. Another measure is the kappa (κ) statistic, which is used to correct for chance agreement. This statistic represents the average rate of agreement for an entire set of scores, and is an analysis of exact agreement with no room for 'close' agreement, making it suitable for nominal and ordinal data, but not continuous data [51].

There are no absolute cut-off points above which an instrument can be said to have acceptable reliability. However, Portney and Watkins [51] suggest guidelines for the interpretation of various measures of reliability (Table 1).

Response stability is usually expressed in terms of the standard error of measurement (SEM) or the coefficient of variation (CV) [51]. The *standard error of measurement* is the standard deviation of the measurement errors and its interpretation is dependent on the type of reliability that is examined. For test-retest reliability, the SEM would indicate the range of scores that may be expected on retesting [51]. If different raters were used, then the SEM would indicate the range of scores that may be expected from another rater. Interpretation of the SEM is based on the normal curve, where there is a 68% chance that a subject's true score would fall within ± 1 SEM of the mean, and a 95% chance of it falling within ± 2 SEM [51].

The *coefficient of variation* is expressed as a percentage and provides a measure of relative variation or response stability across repeated measures [51] or intra-test reliability. In the area of work-related assessments, the coefficient of variation is often used as a measure to determine the consistency or sincerity of effort produced by a subject, rather than as a measure of test stability. The clinician must ensure, therefore, that interpretation of this value is made with regard to the purpose of the calculation of the CV.

When the coefficient of variation is used to determine whether a subject's effort is maximal or submaximal based on the variability of the results there are no clear cutoffs that can be applied [56,57] although there have been some attempts to do this [46]. In terms of instrument precision, coefficients of variation of 6–7% are considered to indicate a high degree of precision, while an instrument with a CV of over 12.5% is considered to have poor precision [60].

1.4. Reliability of work-related assessments

The types of reliability most appropriate for work-related assessments are test-retest/intra-rater and inter-rater reliability. Awareness of the population used as the sample is also extremely important, especially when the purpose of the work-related assessment may be to determine the extent of impairment or disability for compensation benefits.

An example of the importance attributed to reliable and valid assessments in the area of disability determination is seen in the review of functional assessment literature and methods conducted for the United State Social Security Administration (SSA) [53]. The purpose of the review was to

> ". . . thoroughly research the literature about assessment systems, methods, and instruments for associating clinical measures with measures of functional ability and capacity to perform activities and tasks, and develop a systematic method of describing, categorising, comparing, and evaluating them for the purpose of determining their potential application in the disability insurance program" [53, p. 3].

In the SSA review, the criteria for automatic exclusion of an instrument from further review were no evidence of reliability or validity, and no citations of research. It is clear that these aspects were considered crucial before an instrument was even considered for review.

Some commercially available work-related assessments, however, have no reliability studies that were located. Other assessments rely on previous research that examined some sub-tests that are now incorporated into the current battery. Still others are proprietary systems that do not release information other than to purchasers of the systems. This provides a wide range of information that the clinician must be aware of in order to consider the appropriateness or relevance of the work-related assessment to the clinical situation.

Table 1
Interpretation of measures of reliability

Measure of reliability	Range of values	Interpretation of values
Correlation Coefficients (r)	0.00–0.25	Little or no relationship
	0.26–0.50	Poor to fair
	0.51–0.75	Moderate to good
	> 0.75	Good to excellent
	$\geqslant 0.90$	Required for clinical application to ensure valid interpretation of findings
Intra-class Correlation Coefficient (ICC)	$\leqslant 0.75$	Poor to moderate reliability
	> 0.75	Good reliability
	$\geqslant 0.90$	Required for clinical application to ensure valid interpretation of findings
Kappa (κ) value	0.00	Chance agreement
	0.01–0.40	Poor to fair agreement
	0.41–0.60	Moderate agreement
	0.61–0.80	Substantial agreement
	0.81–1.00	Excellent to perfect agreement

2. Method

This study surveyed an extensive range of information sources to determine the extent of evidence of reliability of all forms for work-related assessments. Methods used to access these sources were:

- CD-ROM searches of the CINAHL (1980 – Dec 1997), Medline (1970 – Dec 1997), PsychInfo (1984 – Dec 1997) and ACEL Occupational Health and Safety databases, using the key words 'functional capacity evaluation', 'vocational assessment', 'work assessment', 'work evaluation', 'work sample', and the specific names of the various assessments (e.g., Progressive Isoinertial Lifting Evaluation, Valpar);
- Using secondary sources (i.e., reference lists from published articles) to locate further literature;
- Examining administration and procedure manuals for specific assessments when these were available;
- Contacting distributors of specific assessments;
- Accessing proceedings of conferences where it was known papers had been presented on specific work-related assessments; and
- Accessing theses, or abstracts of theses, where it was known that research had been conducted on specific work-related assessments.

Fifty-five different work-related assessments were identified. Of these, 28 were included in this review. The work-related assessments considered in this review are those that: (1) are currently in use in occupational rehabilitation in Australia, (2) are currently commercially available or still in use, (3) are referred to in publications, and (4) focus predominantly on physical factors associated with work.

The 28 assessments included in this study are: Acceptable Maximum Effort (AME), Applied Rehabilitation Concepts (ARCON), AssessAbility, Blankenship Functional Capacity Evaluation, BTE Work Simulator, California Functional Capacity Protocol (Cal-FCP), Dictionary of Occupational Titles – Residual Functional Capacity (DOT-RFC), EPIC Lift Capacity Test, ERGOS Work Simulator, ErgoScience Physical Work Performance Evaluation (PWPE), Isernhagen Functional Capacity Evaluation, Key Method Functional Capacity Assessment, Lido WorkSET, MESA/System 2000, Progressive Isoinertial Lifting Evaluation (PILE), Polinsky Functional Capacity Assessment, Quantitative Functional Capacity Evaluation (QFCE), Singer/New Concepts Vocational Evaluation System (VES), Smith Physical Capacity Evaluation, Spinal Function Sort, Valpar Component Work Samples, WEST Standard Evaluation, WEST 4/4A, WEST Tool Sort and LLUMC Activity Sort, WorkAbility Mark III, Work Box, and WorkHab Australia.

These assessments cover a wide range of work demands and include instruments that are based on individual self-perception of performance (Spinal Function Sort, WEST Tool & LLUMC Activity Sorts), as well as those reliant on the observation skills of the clinician (e.g., Isernhagen FCE, PWPE, Smith PCE). Some instruments are computerised (ARCON, BTE Work Simulator, ERGOS Work Simulator, Lido WorkSET), while others have specific equipment that is used (e.g., Blankenship FCE, Valpar CWS, WorkAbility Mk III, WorkHab Australia). A number focus specifically on lifting (e.g., EPIC Lift Capacity Test, PILE, WEST Standard Evaluation), while others cover the wide gamut of physical demands (e.g., AssessAbility, Blankenship FCE, Cal-FCP, DOT-RFC, Isernhagen FCE, Polinsky FCA).

There are several assessments that are no longer commercially available (i.e., Lido WorkSET, Polinsky FCA, Singer/New Concepts VES) although they may still be in use by clinicians. For this reason they are included in this study. There are several other work-related assessments, however, that have not been included. These are the FFFWA (Functionally Fit For Work Analysis), referred to by Tramposh [66], and the Physio-Tek and Sweat FCA, both referred to by Lechner et al. [38]. These are the only references to these assessments that were located, and there was no reply to correspondence that was sent to the organisations identified as marketing the products.

Assessments with an emphasis predominantly on clients with developmental disabilities, cognitive deficits or learning disabilities have also been omitted. These are the McCarron-Dial, Micro-TOWER, Philadelphia JEVS (Jewish Employment and Vocational Service), TOWER and Valpar 17 assessments.

Common hand function/dexterity tests have been omitted, as their emphasis is on determining specific aspects of hand function, rather than overall ability for work. Some of these tests, however, are included as sub-tests of assessment batteries. The hand function assessments not examined include the Bennett Hand-Tool Test, Crawford Small Parts Dexterity Test, Grooved Pegboard, Minnesota Dexterity Test, Minnesota Rate of Manipulation Test, O'Connor Finger Dexterity Test, O'Connor Tweezer Dexterity Test, Pennsylvania Bi-Manual Work Sample, Purdue Pegboard and Stromberg Dexterity Test.

Computerised lifting simulators and isokinetic range-of-motion devices have also been omitted. These devices include the Ariel Computerised Exercise (ACE) System Multi-Function Unit, Biodex, Cybex Back Testing System (incorporating the Liftask, Trunk Extension-Flexion and Torso Rotation components), Isostation B-200, Isostation Liftstation, Kin Com, LIDOLift, Lift Trak, Lumbar Motion Monitor, and various other "lifting machines".

2.1. Categorisation of evidence for reliability of work-related assessments

Each work-related assessment included in this study was examined for evidence of reliability and instrument precision. The evidence was categorised according to the quality of the information provided. Each piece of evidence was also critiqued in terms of the study design, subjects, analyses and interpretation of results to enable a judgement to be made on the acceptability of the reliability of the assessment studied. Appendix 1 identifies each of the sources used.

The levels of evidence for the reliability of work-related assessments included in this review were categorised into six broad categories (Table 2). The lowest level (Level 0) indicates that no evidence for reliability was identified. Level 1 indicates that the developers of the assessment relied on previous studies conducted on sub-tests or portions of the assessment. The assumption made by the test developers is that the previous studies demonstrated acceptable reliability and so justifies the inclusion of the sub-test. The danger is generalising acceptable reliability for some sub-tests to all components of the assessment. Furthermore there may be no critical review of the previous studies before accepting the results reported.

Level 2 indicates that although there may be some report of reliability, there is no detail provided to enable the evaluation of results. Level 3 is similar, but some detail is provided to allow a cursory examination of results. Sufficient detail for the evaluation of results consists of a description of the type of reliability studied, the sample used, type of data and how it was collected, analyses used, and interpretation of the results.

Levels 4 and 5 are essentially the same; however, the forum in which the detail and results are presented varies. Both provide sufficient detail for the examination and evaluation of results, with Level 4 reporting these in non-peer-reviewed forums, while Level 5 reports results in peer-reviewed journals.

Some assessments in this study had evidence of reliability from a number of these levels. It should be noted, however, that although the reliability of an assessment has been examined and reported in adequate detail in a peer-reviewed forum (i.e., Level 5), this does not indicate that the reliability is acceptable for clinical purposes.

For each work-related assessment included in this study all available evidence of reliability was located and examined. Following a thorough analysis of the information for the detail necessary to determine the quality and usefulness of the evidence presented, the level of evidence was determined and summarised (see Table 3). For those assessments with acceptable levels of evidence (Levels 4 and 5) the level of reliability was then determined based on the interpretation of measures of reliability described in Table 1 (see Table 4).

3. Results

A summary of the level of evidence for reliability and instrument precision that could be located for the range

Table 2
Levels of evidence for reliability

Level	Description
0	No reliability demonstrated or reported.
1	Reliability is assumed from previous studies conducted on sub-tests now incorporated into the current assessment. Previous studies may be in either a non-peer-reviewed or peer-reviewed forum.
2	Reliability is reported, but there is no detail provided to enable examination of the results. Maybe in either a non-peer-reviewed or peer-reviewed forum.
3	Reliability is reported with some detail to enable a cursory examination of the results, but more detail is required. May be in either a non-peer-reviewed or peer-reviewed forum.
4	Reliability is reported with sufficient detail to enable examination of the results. Results and detail are provided in a non-peer-reviewed forum (i.e., conference presentation, administration manual, book, Honours, Masters or Doctoral thesis).
5	Reliability, with sufficient detail to enable examination of the results, is reported and published in a peer-reviewed forum (i.e., peer-reviewed journal).

of work-related assessments included in this study is presented in Table 3. For those assessments with acceptable levels of evidence (Levels 4 and 5) the level of reliability is reported in Table 4.

3.1. Studies with insufficient evidence for reliability (Levels 0–3)

No reliability studies of any type were identified (Level 0) for the Key FCA, Polinsky FCA, Smith PCE, WEST Tool Sort, or LLUMC Activities Sort (Table 3). It is difficult, therefore, to comment on any aspect of reliability for these assessments.

Some assessments indicated that reliability was acceptable based on the inclusion of various sub-tests that had previously established reliability (Level 1). These included the DOT-RFC (for lifting, carrying, stooping and fingering) [21], the entire QFCE battery [72] and WorkHab Australia (grip strength – [7,48]) (Table 3). It is not possible, however, to comment on the overall reliability for these assessments.

AssessAbility, which is based on MTM (Methods-Time-Measurement) data, is also considered to be at Level 1 as it relies on previous studies using MTM showing "extremely high reliability" [13]. While the use of MTM and other predetermined motion-time standards (PMTS) are considered reliable methods for determining work sample production standards [18,50], there are some limitations that should be noted. PTMS are based on average, experienced workers, which assume that the individual is familiar and proficient with the task being performed [50]. This is not always the case for injured clients being assessed. A second issue is that actual industrial standards are affected by a number of variables and may result in a range of acceptable levels of performance for a job that vary between businesses [50]. While the use of PMTS as a basis for developing an assessment may be acceptable, no formal reliability studies have been reported on AssessAbility.

The Blankenship FCE [5] also incorporates several sub-tests that have been developed by other researchers and have reported reliability (e.g., Oswestry Low Back Pain Disability Questionnaire – [16]; static strength tests using AME – [36]; hand grip strength – [48]). However, no specific reliability studies have been conducted on this assessment battery or portions of it that are unique to the system. None of the assessments with assumed acceptable reliability for various sub-tests and components report the actual results of these prior studies on which acceptance is based, however, references to the studies are provided. Despite several sub-tests having published reliability, the entire Blankenship FCE would be considered to be at Level 1, having only assumed reliability from other sources.

The WorkAbility Mk III has test-retest reliability reported for a number of manual dexterity sub-tests of a previous version of the assessment [55] that was presented in a non-peer-reviewed forum (Level 2). Eleven of 17 sub-tests had good reliability, while only four had coefficients in the excellent range. Correlation coefficients rather than ICCs were calculated and there is insufficient detail regarding the methodology used on which to evaluate the quality of the results.

The MESA (Microcomputer Evaluation & Screening Assessment)/System 2000 has acceptable test-retest reliability reported in the administration manual [6,68], but no other sources of reliability were identified. Minimal information is provided in the manual regarding how the results were obtained (Level 2). A clinician may consider that studies reported in peer-reviewed

Table 3
Summary of level of evidence for reliability of work-related assessments

Assessment	Type of reliability				
	Test-retest/Intra-rater	Inter-rater	Intra-test	Instrument Precision	Other - Alt Forms, Int. Consistency (Split 1/2 & Test Item)
AME	0	0	5 (indicated as T-RT in study)	0	0
ARCON	5	0	0	0	0
AssessAbility	1 (MTM)	0	0	0	0
Blankenship FCE	1 (some sub-tests)	0	1 (some sub-tests)	0	0
BTE Work Simulator	5 (# unknown) 5 (#162) 5 (#161, 701) 4, 5 (#302, 802) 5 (#181, 701, 901)	0	2 (#302, 502, 503, 701) 4, 5 (#302, 802) 5 (#302, 502, 503, 601, 701) 5 (#162)	4 (static & dynamic) 4, 5, 5 (dynamic) 5 (static)	0
Cal-FCP (includes EPIC & SFS)	(See EPIC & SFS)	(See EPIC)	0	0	5 (Test item) (See SFS)
DOT-RFC	1 (lift, carry, stoop, finger)	0	0	0	0
EPIC (PLC II was precursor of EPIC)	2, 5, 5 5 (PLC II)	5	0	0	0
ERGOS Work Simulator	5 (comp & human instructions)	0	5 (comp & human instructions)	0	0
Isernhagen FCE	3 (LMH lifts) 5 (floor-waist lift)	3 (LMH lifts) 4 (lifts) 5 (floor-waist lift)	0	0	0
Key FCA	0	0	0	0	0
Lido WorkSET	5 (#19, 22, 52) 5, 5 (#51)	5 (#51)	5 (#19, 22, 52)	5 (accuracy)	0
MESA/System 2000	2	0	0	0	0
PILE	5 (lumbar & cervical lifts), 5	0	0	0	0
Polinsky FCA	0	0	0	0	0
PWPE	5 (lifts)	3, 5 (all) 5 (lifts)	0	0	0
QFCE	1 (all)	0	0	0	0
Singer/New Concepts VES	5	0	0	0	0
Smith PCE	0	0	0	0	0
Spinal Function Sort	5, 5	0	0	0	5 (Split 1/2), 5 (Test item)
Valpar CWS	4 (#19) 4 (#4)	4 (#19)	0	0	0
WEST Std Evaluation	4 5 (shoulder-eye lift)	4, 4 (MHRWS)	0	4 (weights)	0
WEST 4/4A	0	0	2, 3, 5	0	0
WEST Tool & LLUMC Activities Sorts	0	0	0	0	0
WorkAbility Mk 3	2 (manual dexterity tests)	0	0	0	0
Work Box	5, 5	0	0	0	0
WorkHab	1 (grip strength)	0	0	0	0

N.B. Unless otherwise indicated, the entire assessment was studied. For all other assessments the sub-test or portion of the assessment studied is in parentheses. The items for the BTE Work Simulator, Lido WorkSET and the Valpar Component Work Samples indicate the number of the specific attachment or work sample studied.

Table 4
Summary of level of reliability of work-related assessments

Assessment	Type of reliability				
	Test-retest/Intra-rater	Inter-rater	Intra-test	Instrument Precision	Other - Alt Forms, Int. Consistency (Split 1/2 & Test Item)
AME	Unknown	Unknown	Excellent, with Clinical utility	Unknown	Unknown
ARCON	Fair - Moderate (using revised method)	Unknown	Unknown	Unknown	Unknown
AssessAbility	Unknown	Unknown	Unknown	Unknown	Unknown
Blankenship FCE	Unknown	Unknown	Unknown	Unknown	Unknown
BTE Work Simulator	Good - Excellent, with Clinical utility (#161, 162, 181, 302, 701, 802, 901)	Unknown	Excellent, with Clinical utility (#162, 302, 802); Unable to determine acceptability of CV (#302, 502, 503, 601, 701)	*Static:* Excellent, with Clinical utility *Dynamic:* Poor	Unknown
Cal-FCP (includes EPIC & SFS)	(See EPIC & SFS)	(See EPIC)	Unknown	Unknown	Moderate - Good (Test-item)
DOT-RFC	Unknown	Unknown	Unknown	Unknown	Unknown
EPIC (PLC II was precursor of EPIC)	Good - Excellent, with Clinical utility	Good - Excellent, with Clinical utility	Unknown	Unknown	Unknown
ERGOS Work Simulator	Good - Excellent, with clinical utility (comp & human instructions)	Unknown	Unable to determine acceptability of CV (comp & human instructions)	Unknown	Unknown
Isernhagen FCE	Substantial (floor-waist lift)	Substantial (floor-waist lift); Unable to determine (lifts)	Unknown	Unknown	Unknown
Key FCA	Unknown	Unknown	Unknown	Unknown	Unknown
Lido WorkSET	Good - Excellent, with Clinical utility (#19, 22, 51, 52)	Excellent, with Clinical utility (#51)	Unable to determine acceptability of CV (#19, 22, 52)	Excellent (torque); Acceptable (work & power)	Unknown
MESA/System 2000	Unknown	Unknown	Unknown	Unknown	Unknown
PILE	Good - Excellent, with Clinical utility	Unknown	Unknown	Unknown	Unknown
Polinsky FCA	Unknown	Unknown	Unknown	Unknown	Unknown
PWPE	Excellent, with Clinical utility (lifts)	Moderate - Excellent	Unknown	Unknown	Unknown
QFCE	Unknown	Unknown	Unknown	Unknown	Unknown
Singer/New Concepts VES	Moderate	Unknown	Unknown	Unknown	Unknown
Smith PCE	Unknown	Unknown	Unknown	Unknown	Unknown
Spinal Function Sort	Good - Excellent	Unknown	Unknown	Unknown	Excellent, with Clinical utility (Split 1/2 & Test-item)
Valpar CWS	Moderate - Excellent (#19) Moderate - Good (#4)	Substantial - Excellent, with Clinical utility (#19)	Unknown	Unknown	Unknown
WEST Std Evaluation	Unable to determine acceptability	Poor - Excellent (MHRWS)	Unknown	Unacceptable (weights)	Unknown

Table 4 continued

Assessment	Type of reliability				
	Test-retest/Intra-rater	Inter-rater	Intra-test	Instrument Precision	Other - Alt Forms, Int. Consistency (Split 1/2 & Test Item)
WEST 4/4A	Unknown	Unknown	Unable to determine acceptability of CV	Unknown	Unknown
WEST Tool & LLUMC Activities Sorts	Unknown	Unknown	Unknown	Unknown	Unknown
WorkAbility Mk 3	Unknown	Unknown	Unknown	Unknown	Unknown
Work Box	Moderate - Good (overall); Excellent, with clinical utility (F); Poor (M)	Unknown	Unknown	Unknown	Unknown
WorkHab	Unknown	Unknown	Unknown	Unknown	Unknown

N.B. The assessments in **bold** are those with evidence at Level 4 or 5. The sub-test or portion of the assessment studied is in parentheses. The items for the BTE Work Simulator, Lido WorkSET and the Valpar Component Work Samples indicate the number of the specific attachment or work sample studied.

forums are required to support these findings.

The work-related assessments with little evidence of reliability (Levels 0–3) may have good or poor reliability. Without adequate evidence, however, it is impossible for clinicians to determine whether an assessment's reliability is of an acceptable level for their purposes.

3.2. *Studies with sufficient evidence for reliability (Levels 4–5)*

The Valpar Component Work Samples reported test-retest reliability for each work sample in the original administration manuals (1974), however, Botterbusch [6] indicated that the data available could not be assessed to determine the acceptability or otherwise of the reported reliability. This reliability information is no longer included in the revised manuals for these work samples for this reason [10]. There is, however, some evidence of moderate to good reliability for two of the work samples (VCWS 4 and 19), albeit from non-peer-reviewed sources [3,67] (Table 4). There are, however, sufficient details on which to critically evaluate the findings (Level 4).

The AME (Acceptable Maximum Effort) assessment [36] identifies the type of reliability examined to be test-retest, however, with retesting occurring within 30 to 60 seconds, it would appear to be more accurately described as intra-test reliability. The results indicate that the assessment is highly stable, with all correlation coefficients over 0.92. There is no indication, however, whether this stability is maintained over a longer period of time than 1 minute.

The ARCON (Applied Rehabilitation Concepts) also has only one study examining test-retest reliability [25] (Table 3). It identifies unacceptable levels of reliability and attempts to improve them through modifications to stabilisation of the individual and improved sensor placement. The results, although improved, continue to indicate unacceptable test-retest reliability with only three of ten sub-tests for males or females having correlation coefficients over 0.70. Correlation coefficients, rather than ICCs were used to determine reliability.

The Singer/New Concepts VES (Vocational Evaluation System) was reported to have moderate test-retest reliability with changes in retesting attributed to improvement from a training program [11]. One would question the appropriateness of examining test-retest reliability in a situation where it is known that the subjects are involved in an intervention program, and change is in fact desired.

The WEST 4/4A had the coefficient of variation for males and females reported [46,71] with no indication of how the results had been obtained for males (Level 2), and some basic information for some of the female data (Level 3). As there is no indication of the original source of data reported by Matheson and Ogden-Niemeyer [46], it is difficult to interpret the results presented. The reported 12% CV could be interpreted as an unacceptably unstable instrument (Solgaard et al. [60] considered a grip strength device with a CV of 12.7% to be unacceptable). A mean CV of 12% for supination and 10.4% for pronation was found by Innes et al. [30] (Level 5). This study also found CVs greater than 12% in 50% of subjects for supination

and 27% for pronation, causing concern about the usefulness of this instrument due to its instability. Without clear guidelines for the interpretation of CV, however, it is difficult to determine the acceptability or otherwise of this result.

The Work Box has effectively used the examination of test-retest reliability to refine and improve administration procedures [4,61]. It also has identified those clients for whom the assessment will most clearly indicate a change in performance [61], however, there was no demonstration of actual change occurring. A difference in test-retest reliability between genders has also been reported [61], with excellent reliability in males, and poor reliability in females (Table 4), indicating that the assessment is more appropriately used with males.

The BTE Work Simulator appears to be the most thoroughly researched instrument, with test-retest, intra-test and instrument reliability all being investigated on a number of occasions with subjects from healthy and injured populations. It should be noted, however, that the BTE has 22 different attachments and operates in both the static and dynamic modes. The static mode is highly reliable [2,19,20], however, there is concern regarding reliability of the dynamic mode [8, 9,12,15,19,20]. Less than a third of the attachments have been examined for test-retest and/or intra-test reliability, and there are no studies of inter-rater reliability. All test-retest studies in the static mode indicate good to excellent reliability. It should be noted, however, that some studies used correlation coefficients rather than the preferred ICC to calculate reliability (e.g., [2, 35].

Other work simulators include the ERGOS and Lido WorkSET, both more recent additions than the BTE. The Lido WorkSET has multiple attachments, similar to the BTE. Four attachments have been used to examine test-retest, three for intra-test and one for inter-rater reliability. Results indicate that there is a good to excellent level of reliability for the attachments studied in static and dynamic modes (Table 4). Instrument precision has also been examined by using calibrated weights to generate torque [43]. It was found to be excellent for torque measurement, while work and power measurements were acceptable.

The ERGOS Work Simulator has only been examined for the reliability of responses using the static lifting sub-tests and varying methods of delivery of instruction (i.e., computer versus human instructions) [42]. The results indicate good reliability for both forms of instruction with no significant difference between them.

The Cal-FCP includes the EPIC Lift Capacity Test and the Spinal Function Sort, amongst its components [45]. This is the only physical test battery that has had internal consistency, in this case test item reliability, examined (Table 3). Variables or sub-tests expected to be related (e.g., pinch strength and grip strength; Spinal Function Sort and lift capacity) were examined. Regression equations were found to be highly significant.

The Spinal Function Sort included in the Cal-FCP but able to be used alone, is used to determine perceived, rather than actual, capacity or performance. Acceptable test-retest reliability has been found, as has internal consistency (both split-half and test item) [22,44] (Table 3). These reliability results were from subjects reporting back pain.

The EPIC Lift Capacity Test, also included in the Cal-FCP and able to be used independently, has been examined for both test-retest and inter-rater reliability. Both forms were at good to excellent levels, with test-retest having been established on a number of different occasions in a variety of settings with large numbers of subjects, strengthening the findings (Table 3).

The PILE (Progressive Isoinertial Lifting Evaluation) and the WEST Standard Evaluation are also assessments of lifting capacity. Only test-retest reliability has been examined for the PILE, and while the results are acceptable, the subject numbers on which this is based are small ($n = 10$) [49].

The WEST Standard Evaluation's test-retest reliability is not clearly established. Matheson's study [41] only examined the reliability of the load lifted through a limited range (shoulder height to eye level) and reported the coefficient of variation based on three lifting trials (Level 5). No other statistical analysis was undertaken. In a non-peer-reviewed study by Tan [64, 65] test-retest reliability of the maximum weight lifted was determined using t-tests, rather than statistics such as the ICC (Level 4). Both Matheson [41] and Tan [64, 65] consider that acceptable test-retest reliability has been demonstrated, however, this is questionable based on the type of analyses performed (Table 4).

Inter-rater reliability for the WEST Standard Evaluation has only been reported in non-peer-reviewed forums [26,54,64,65] (Level 4). Using a large number of raters ($n = 18$), Hehir [26,54] reported an overall rating of fair reliability for the determination of a safe lifting technique, while Tan [64,65] used only two raters with a large number of clients and reported higher levels of agreement.

Instrument precision in the form of the accuracy of the weights used in the assessment was examined by

Tan [64] (Level 4). For one set of equipment and weights it was found that all were under the reported weights. This indicates the need to check equipment and certainly raises concerns, but the results cannot be generalised beyond the study.

The Isernhagen FCE has only had the lifting component of this test battery examined. The ability to determine a safe lift and/or rate it as light, medium or heavy was the focus of the inter-rater studies (Table 3). Smith [59] found substantial agreement between raters, however, raters were only asked to determine if a lift was safe or unsafe (Level 5). Similar results were obtained by Isernhagen and Hart (cited in [33] (Level 3). A third inter-rater study [17] used inappropriate statistics (t-tests to determine reliability between four raters), so it is not possible to determine if the results were at an acceptable level (Level 4).

The test-retest studies of the Isernhagen FCE were based on showing the same video of persons lifting to a group of raters on two separate occasions. In the case of the Smith study, the two viewings were one day apart. It is questionable whether this methodology is an acceptable way to determine test-retest reliability of an assessment.

The ErgoScience PWPE (Physical Work Performance Evaluation) is one of the more recent assessments available. As appropriate for an assessment reliant on the judgement of the evaluator, inter-rater reliability was examined and found to have moderate to excellent agreement for the major sections and overall [39,40] (Levels 3 and 5). These findings for determining maximum lifting capacity were supported by a pilot study [58] that also found high levels of test-retest reliability for a small group (Level 5).

4. Discussion

4.1. Evidence of reliability

The results of this study indicate that the evidence for reliability of a wide range of work-related assessments ranges from non-existent to being investigated and reported in sufficient detail to enable decisions to be made regarding their clinical utility. There does not appear to be a single assessment that has been thoroughly and comprehensively investigated for all relevant aspects of reliability. Some systems, however, appear to have a promising basis on which to build further studies to provide sufficient evidence of reliability.

The work-related assessments where there was no evidence for reliability (Level 0) were the Key FCA, Polinsky FCA, Smith PCE, and WEST Tool and LLUMC Activities Sorts. The first three systems assess whole body function and physical demands, while the latter two address self-perception of functional ability. These assessments were developed in the 1970s and early 1980s and include proprietary systems (Key FCA, Polinsky FCA) that release limited information to those not trained in their use. This lack of information makes it extremely difficult for clinicians to determine if these assessments are acceptable in terms of reliability particularly before they purchase the systems.

A number of assessments include test components that have been previously developed and investigated. These are AssessAbility, Blankenship FCE, DOT-RFC, QFCE, and WorkHab Australia. For some assessments it may be only a single previously developed sub-test, while others make up the entire test battery. Reliability is assumed to be acceptable from the results of the previous studies (Level 1).

All the assessments with Level 1 evidence for reliability are whole-body, physical demand assessment batteries made up of numerous sub-tests. With extensive assessment batteries it is not surprising that previously developed components are incorporated to avoid "reinventing the wheel". The concern, however, is the reliance on previous studies as the only form of reliability, rather than the actual results of the previous studies. Acceptable reliability for sub-tests cannot be extrapolated to the entire assessment battery. The interaction of these sub-tests with other components of the battery may also impact on overall reliability.

MESA/System 2000 and WorkAbility Mk III have reports of reliability studies, however, there is insufficient detail to enable a thorough examination of the results (Level 2). Both of these assessments are very different from the other, with the only similarity being when they were originally developed. Both were developed in the 1970s and early 1980s, although System 2000 and WorkAbility Mk III are updates of the original assessments.

The WEST 4/4A has evidence at Levels 2, 3 and 5, however, this is only for intra-test reliability reported as a coefficient of variation (CV). Without clear guidelines for interpreting the CV it is not possible to determine the clinical acceptability of the results. However, there may be some question regarding the usefulness of an instrument with a CV of 12% [60].

The Isernhagen FCE has some studies reporting reliability with only cursory detail (Level 3). This evidence

is in the form of conference abstracts or summaries that only provide very brief or sketchy details, which make it difficult to thoroughly evaluate the results that are presented. The study by Farag [17] (Level 4) used inappropriate statistics and so cannot be used to determine reliability, however, Smith [59] found substantial agreement between raters when determining the safety of a floor to waist lift.

There is limited evidence at Level 4 where the reliability is reported in sufficient detail to enable a thorough examination and is reported in a non-peer-reviewed forum. Only studies investigating the Isernhagen FCE (lifting), Valpar Component Work Samples 4 and 19, and WEST Standard Evaluation were identified. These studies were all conducted in Australia.

It is recognised that a limitation of this study is that evidence of reliability at Level 4 may not have been located, as reference to these studies is very limited and obtaining them is equally difficult. It is possible that there are many more studies at this level, but they were not located for this study. This limitation highlights the importance of researchers at all levels to publish their findings in public forums that are accessible around the world rather than in a limited geographical region.

A positive outcome of this study was the number of work-related assessments found to have evidence of reliability reported in sufficient detail to enable evaluation and presented in a peer-reviewed forum (Level 5). AME, ARCON, BTE Work Simulator, Cal-FCP, EPIC Lift Capacity, ERGOS Work Simulator, Isernhagen FCE (Lifting), Lido Work Set, Spinal Function Sort, PILE, PWPE, Singer VES, WEST Standard Evaluation and the Work Box all have reliability studies published in peer-reviewed journals. It should be noted, however, that while it is commendable that these studies are published, it does not indicated that these assessments have acceptable reliability for clinical purposes. Also, many have only investigated some portions of the overall assessment for reliability, or only have some forms of reliability investigated.

A further concern is the potential for bias against the publication of studies reporting poor reliability results. This may be an issue when the developers of an assessment are also those who conduct the research (e.g., Fishbain et al. for the DOT-RFC; Isernhagen for the Isernhagen FCE; Khalil et al for the AME; Lechner for the PWPE; Matheson for the Cal-FCP, EPIC, Spinal Function Sort, WEST Standard Evaluation and WEST 4/4A; Mayer et al for the PILE; Shervington for WorkAbility Mk 3). When these studies are published in a peer-reviewed forum there is an opportunity for independent reviewers to comment on the study and the interpretation of the findings. This does not, however, address the issue that studies demonstrating poor levels of reliability are not submitted for publication in the first place. There is, however, no suggestion that this has occurred for the assessments reviewed in this study.

Perhaps of greater concern is when independent studies are conducted and the manufacturer of the instrument attempts to block the dissemination of the results. An example of this was reported by Strong and Westmorland [62] where studies demonstrating poor reliability for the dynamic mode of the BTE work simulator resulted in an unsuccessful law suit against the researcher and an attempt to withhold publication of another study, although it was subsequently published.

4.2. Level of reliability

For assessments with an adequate level of evidence (i.e., Level 4 or 5), it is possible to also comment on the acceptability of the reliability for clinical purposes using the guidelines previously described (see Table 1). It is not possible to comment on the reported reliability for those assessments with evidence at lower levels (0–3) due to the lack of necessary information to enable an adequate critique of the results presented.

The assessment with a consistently good level of instrument precision, test-retest and intra-test reliability across a number of attachments is the BTE Work Simulator. This is only true, however, when it is used in the static mode. Similarly high levels of reliability have been demonstrated by the Lido WorkSET, however, this has been on a fewer number of attachments. Both systems use a computer-based system to determine results.

Those assessments that focus on a single or limited number of aspects of work-related function, such as the EPIC Lift Capacity, PILE, Valpar Component Work Samples (#4 and #19) and the Work Box, also demonstrate good levels of reliability. The EPIC in particular has consistently demonstrated good to excellent reliability over a number of studies, therefore increasing confidence in this assessment.

Work-related assessments consisting of multiple sub-tests or components range from having no evidence of reliability at all (Key FCA, Polinsky FCA, Smith PCE), through those that assume reliability from previously developed sub-tests (AssessAbility, Blankenship FCE, DOT-RFC, QFCE, WorkHab), to those that have examined at least some components for reliability (Isernhagen FCE, PWPE, WorkAbility Mk 3). It is not possible to comment on the level of reliability for the

first two groups. Of the latter group, the lifting components of the Isernhagen and PWPE have acceptable test-retest reliability. The PWPE, however, has also demonstrated acceptable inter-rater reliability for the entire assessment, while the Isernhagen FCE has only focussed on the lifting component. WorkAbility Mk 3 has some promising results for its manual dexterity sub-tests, however, further detail in a peer-reviewed forum is required before more confidence can be placed in this assessment. All require further and ongoing studies to cover all aspects of the assessments.

4.3. Reliability versus validity

While the emphasis of this study has been to examine the evidence of reliability for a range of work-related assessments, it is also necessary to consider the usefulness of these results. It is commonly viewed that a test which has poor reliability, cannot be said to be valid, and therefore should not be used to make clinical decisions [24]. It is of interest, therefore, to note comments regarding the compromise which is made between reliability and validity when attempting to simulate the actual work environment or requirements "as the test situation simulates reality more closely, control becomes more difficult. . . the more closely one tries to simulate a real criterion situation, the less reliable will be one's measurement of performance" (Fitzpatrick & Morrison, 1971, p. 240, cited in [52, p. 46]). This has major implications for assessments, such as workplace-based or situational assessments, which attempt to increase face validity by using the actual workplace environment and/or tasks. They could, however, be considered to have potentially poor reliability, due to the lack of control the evaluator has over the variables that may affect performance.

Standardised instruments that are administered in controlled environments have a greater chance of having acceptable consistency or reliability. These types of instruments tend to be those which assess performance at the skill or task level, rather than at the activity or role level as may be done with a workplace-based assessment. Differentiation between these levels of performance has previously been described in detail [31].

5. Conclusion

There remain many assessments developed in the 1970s and 1980s that do not have sufficient evidence of their reliability. It appears that the use of these assessments continues without question "because they are there", and historically there were no other options. Developers of work-related assessments in the 1990s appear to have a greater appreciation and understanding of the need to investigate and report the reliability of the assessments used by clinicians in a variety of contexts and for different purposes. Others, including clinicians and academics, also recognise this need and are conducting independent research into these instruments.

For the limited number of work-related assessments with an adequate level of evidence on which to judge their reliability, most demonstrate a moderate to good level. Few work-related assessments, however, have demonstrated levels of reliability sufficient for clinical (and legal) purposes. Fewer still have this demonstrated over a number of studies, in varying contexts and with different populations.

While current researchers of work-related assessments are to be commended for their concern regarding the need to demonstrate the reliability of the instruments used, it is also essential that this research continue. A single study of only one form of reliability of a portion of an assessment battery is insufficient for clinical purposes if therapists are to have confidence in the tools they are using. Ongoing research in this area is a necessity, particularly with a diverse range of injured as well as uninjured populations.

With this review clinicians are now able to examine their options with regard to the reliability of the assessments they choose to use. Interpretation of changes in test results can now be considered in the light of the evidence for the reliability of the instrument used.

References

[1] Abdel-Moty, E., Compton, R., Steele-Rosomoff, R., Rosomoff, H. and Khalil, T.M., Process analysis of functional capacity assessment, *Journal of Back & Musculoskeletal Rehabilitation* **6** (1996), 223–236.

[2] Anderson, P.A., Chanoski, C.E., Devan, D.L., McMahon, B.L. and Whelan, E.P., Normative study of grip and wrist flexion strength employing a BTE work simulator, *Journal of Hand Surgery* **15A**(3) (1990), 420–425.

[3] Barrett, T., Browne, D., Lamers, M. and Steding, E., Reliability and validity testing of Valpar 19, Perth, WA.: AAOT, *Proceedings of the 19th National Conference of the Australian Association of Occupational Therapists* **2** (1997), 179–183.

[4] Black, M.K., Nelson, C.E., Maurer, P.A. and Bauer, D.F., Test-retest reliability of the Work Box: A work sample with standard instructions, *Work* **3**(4) (1993), 26–34.

[5] Blankenship, K.L., *The Blankenship system functional capacity evaluation: The procedure manual*, (2nd ed.), Macon, GA: The Blankenship Corporation, 1994.

[6] Botterbusch, K.F., *Vocational assessment and evaluation systems: A comparison*, Menomonie, Wisconsin: Materials Development Center, Stout Vocational Rehabilitation Institute, University of Wisconsin-Stout, 1987.
[7] Bradbury, S. and Roberts, D., *WorkHab Australia Functional Capacity Evaluation workshop manual*, Bundaberg, Qld: WorkHab Australia, 1996.
[8] Cetinok, E., Coleman, E., Fess, E.E., Dunipace, K.R. and Renfro, R., Reliability of the BTE work simulator dynamic mode [Abstract], *Journal of Hand Therapy* **8**(1) (1995a), 52–53.
[9] Cetinok, E.M., Renfro, R.R. and Coleman, E.F., A pilot study of the reliability of the dynamic mode of one BTE work simulator, *Journal of Hand Therapy* **8**(3) (1995b), 199–205.
[10] Christopherson, B.B., *Revision of manuals for the Valpar Component Work Sample series*, [Publication of limited circulation available from Valpar International Corporation, Tucson, Arizona.] 1991.
[11] Cohen, C. and Drugo, J., Test-retest reliability of the Singer vocational evaluation system, *Vocational Guidance Quarterly* **24**(3) (1976), 267–270.
[12] Coleman, E.F., Renfro, R.R., Cetinok, E.M., Fess, E.E., Shaar, C.J. and Dunipace, K.R., Reliability of the manual dynamic mode of the Baltimore Therapeutic Equipment Work Simulator, *Journal of Hand Therapy* **9**(3) (1996), 223–237.
[13] Coupland, M., *AssessAbility manual*, Austin, TX: IME AssessAbility Inc., 1995.
[14] Dane, F.C., *Research methods*, Pacific Grove, CA: Brooks/Cole Publishing, 1990.
[15] Dunipace, K.R., Reliability of the BTE work simulator dynamic mode [Letter to the editor], *Journal of Hand Therapy* **8**(1) (1995), 42–43.
[16] Fairbank, J.C.T., Couper, J., Davies, J.B. and O'Brien, J.P., The Oswestry Low Back Pain Disability Questionnaire, *Physiotherapy* **66**(8) (1980), 271–273.
[17] Farag, I., *Functional assessment approaches*, Unpublished Master of Safety Science thesis, University of New South Wales, Kensington, NSW, 1995.
[18] Farrell, J.M., Predetermined motion-time standards in rehabilitation: A review, *Work* **3**(2) (1993), 56–72.
[19] Fess, E.E., Correction: Instrument reliability of the BTE Work Simulator: A preliminary study, *Journal of Hand Therapy* **6**(2) (1993a), 82.
[20] Fess, E.E., Instrument reliability of the BTE work simulator: A preliminary study [Abstract], *Journal of Hand Therapy* **6**(1) (1993b), 59–60.
[21] Fishbain, D.A., Abdel-Moty, E., Cutler, R., Khalil, T.M., Sadek, S., Rosomoff, R.S. and Rosomoff, H.L., Measuring residual functional capacity in chronic low back pain patients based on the Dictionary of Occupational Titles, *Spine* **19**(8) (1994), 872–880.
[22] Gibson, L. and Strong, J., The reliability and validity of a measure of perceived functional capacity for work in chronic back pain, *Journal of Occupational Rehabilitation* **6**(3) (1996), 159–175.
[23] Gibson, L. and Strong, J., A review of functional capacity evaluation practice, *Work* **9**(1) (1997), 3–11.
[24] Hart, D.L., Tests and measurements in returning injured workers to work, in: *The comprehensive guide to work injury management*, S.J. Isernhagen, Ed., Gaithersburg, MD: Aspen, 1995, pp. 345–367.
[25] Hasten, D.L., Johnston, F.A. and Lea, R.D., Validity of the Applied Rehabilitation Concepts (ARCON) system for lumbar range of motion, *Spine* **20**(11) (1995), 1279–1283.
[26] Hehir, A., *A study of interrater agreement and accuracy of the WEST Standard Evaluation*, Unpublished Honours thesis, School of Occupational Therapy, The University of Sydney, 1995.
[27] Innes, E., *Work evaluation systems - What are our current options?* Paper presented at the 6th State Conference of the NSWAOT, Mudgee, NSW, 1993, October.
[28] Innes, E., Workplace-based occupational rehabilitation in New South Wales, Australia, *Work* **5**(2) (1995), 147–152.
[29] Innes, E., Work assessment options and the selection of suitable duties: An Australian perspective, *New Zealand Journal of Occupational Therapy* **48**(1) (1997), 14–20.
[30] Innes, E., Hargans, K., Turner, R. and Tse, D., Torque strength measurements: An examination of the interchangeability of results in two evaluation devices, *Australian Occupational Therapy Journal* **40**(3) (1993), 103–111.
[31] Innes, E. and Straker, L., A clinician's guide to work-related assessments: 2 - Design problems, *Work* **11**(2) (1998a), 191–206.
[32] Innes, E. and Straker, L., A clinician's guide to work-related assessments: 3 - Administration and interpretation problems, *Work* **11**(2) (1998b), 207–219.
[33] Isernhagen Work Systems, *Reliability and validity of the Isernhagen Work systems Functional Capacity Evaluation*, [Publication of limited circulation available from Isernhagen Work Systems, Duluth, Illinois.] 1996.
[34] Johnson, L.J., The kinesiophysical approach matches worker and employer needs, in: *The comprehensive guide to work injury management*, S.J. Isernhagen, Ed., Gaithersburg, MD: Aspen, 1995, pp. 399–409.
[35] Kennedy, L.E. and Bhambhani, Y.N., The Baltimore Therapeutic Equipment work simulator: Reliability and validity at three work intensities, *Archives of Physical Medicine & Rehabilitation* **72** (1991), 511–516.
[36] Khalil, T.M., Goldberg, M.L., Asfour, S.S., Moty, E.A., Rosomoff, R.S. and Rosomoff, H.L., Acceptable maximum effort (AME): A psychophysical measure of strength in back pain patients, *Spine* **12**(4) (1987), 372–376.
[37] Krefting, L.M. and Bremner, A., Work evaluation: Choosing a commercial system, *Canadian Journal of Occupational Therapy* **52**(1) (1985), 20–24.
[38] Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.
[39] Lechner, D.E., Roth, D.L., Jackson, J.R. and Straaton, K., Interrater reliability and validity of a newly developed FCE: The physical work performance evaluation [Abstract], *Physical Therapy* **73**(6 Suppl) (1993), S27.
[40] Lechner, D.E., Jackson, J.R., Roth, D.L. and Straaton, K.V., Reliability and validity of a newly developed test of physical work performance, *Journal of Occupational Medicine* **36**(9) (1994), 997–1004.
[41] Matheson, L.N., Evaluation of lifting and lowering capacity, *Vocational Evaluation & Work Adjustment Bulletin* **19**(3) (1986), 107–111.
[42] Matheson, L.N., Danner, R., Grant, J. and Mooney, V., Effect of computerised instructions on measurement of lift capacity: Safety, reliability, and validity, *Journal of Occupational Rehabilitation* **3**(2) (1993a), 65–81.
[43] Matheson, L.N., Mangesh, G., Segal, J.H., Grant, J.E., Comisso, K. and Westing, S., Validity and reliability of a new device to simulate upper extremity work demands, *Journal of Occupational Rehabilitation* **2**(3) (1992), 109–122.

[44] Matheson, L.N., Matheson, M.L. and Grant, J., Development of a measure of perceived functional ability, *Journal of Occupational Rehabilitation* **3**(1) (1993b), 15–30.
[45] Matheson, L.N., Mooney, V., Grant, J.E., Leggett, S. and Kenny, K., Standardised evaluation of work capacity, *Journal of Back & Musculoskeletal Rehabilitation* **6** (1996), 249–264.
[46] Matheson, L.N. and Ogden-Niemeyer, L., *Work capacity evaluation: Systematic approach to industrial rehabilitation*, (Revised ed.), Anaheim, CA: ERIC, 1987.
[47] Mathiowetz, V., Role of physical performance component evaluations in occupational therapy functional assessment, *American Journal of Occupational Therapy* **47**(3) (1993), 225–230.
[48] Mathiowetz, V., Weber, K., Volland, G. and Kashman, N., Reliability and validity of grip and pinch strength evaluations, *Journal of Hand Surgery* **9A**(2) (1984), 222–226.
[49] Mayer, T.O., Barnes, D., Kishino, N.O., Nichols, G., Gatchel, R.J., Mayer, H. and Mooney, V., Progressive isoinertial lifting evaluation I: A standardised protocol and nor mative database, *Spine* **13**(9) (1988), 993–997.
[50] McCray, P., Competitive work sample norms and standards: Some considerations, *Vocational Evaluation & Work Adjustment Bulletin* **12**(3) (1979), 24–26.
[51] Portney, L.G. and Watkins, M.P., *Foundations of clinical research: Applications to practice*, Norwalk, Connecticut: Appleton & Lange, 1993.
[52] Rosen, G.A., The problem and utility of work sample reliability data, *Vocational Evaluation & Work Adjustment Bulletin* **11**(3) (1978), 45–50.
[53] Rucker, K.S., Wehman, P. and Kregel, J., *Analysis of functional assessment instruments for disability/rehabilitation programs* (Summary report SSA Contract No. 600-95-21914). Richmond, VA: Virginia Commonwealth University, 1996.
[54] Ryan, A., An interrater agreement and accuracy study on the WEST Standard Evaluation [Abstract], *Australian Occupational Therapy Journal* **43**(3/4) (1996), 185.
[55] Shervington, J., *Workplace capability assessment*, Paper presented at the Australia & New Zealand MODAPTS Association International Conference, Melbourne, Vic., 1990, March.
[56] Simonsen, J.C., Coefficient of variation as a measure of subject effort, *Archives of Physical Medicine & Rehabilitation* **76**(6) (1995), 516–520.
[57] Simonsen, J.C., Validation of sincerity of effort, *Journal of Back & Musculoskeletal Rehabilitation* **6** (1996), 289–295.
[58] Smith, E.B., Rasmussen, A.A., Lechner, D.E., Gossman, M.R., Quintana, J.B. and Grubbs, B.L., The effects of lumbosacral support belts and abdominal muscle strength on functional lifting ability in healthy women, *Spine* **21**(3) (1996), 356–366.
[59] Smith, R.L., Therapists' ability to identify safe maximum lifting in low back pain patients during functional capacity evaluation, *Journal of Orthopaedic & Sports Physical Therapy* **19**(5) (1994), 277–281.
[60] Solgaard, S., Kristiansen, B. and Jensen, J.S., Evaluation of instruments for measuring grip strength, *Acta Orthopaedica Scandinavica* **55**(5) (1984), 569–572.
[61] Speller, L., Trollinger, J.A., Maurer, P.A., Nelson, C.E. and Bauer, D.E., Comparison of the test-retest reliability of the Work Box using three administrative methods, *American Journal of Occupational Therapy* **51**(7) (1997), 516–522.
[62] Strong, S. and Westmorland, M., *Determining claimant effort and maximum voluntary effort testing: A discussion paper* (Report), Hamilton, Ontario: Work Function Unit, McMaster University, 1996.
[63] Tan, H.L., *Investigation of the concurrent validity of an assessment component of the WEST Standard Evaluation for use within Australian population and the accuracy of the WEST 3 Comprehensive Weight System*, Unpublished Honours thesis, School of Occupational Therapy, Faculty of Health Sciences, Curtin University of Technology, Perth, W.A., 1995.
[64] Tan, H.L., *Study of the inter-rater, test-retest reliability and content validity of the WEST Standard Evaluation*, Unpublished Masters thesis, School of Occupational Therapy, Faculty of Health Sciences, Curtin University of Technology, Perth, WA., 1996.
[65] Tan, H.L., Barrett, T. and Fowler, B., Study of the inter-rater, test-retest reliability and content validity of the WEST Standard Evaluation, Perth, WA: AAOT, *Proceedings of the 19th National Conference of the Australian Association of Occupational Therapists* **2** 1997, 245–251.
[66] Tramposh, A.K., The functional capacity evaluation: Measuring maximal work abilities, *Occupational Medicine: State of the Art Reviews* **7**(1) (1992), 113–124.
[67] Trevitt, N., *A test-retest reliability study on the Valpar Component Work Sample 4*, Unpublished Honours thesis, School of Occupational Therapy, Faculty of Health Sciences, University of Sydney, Sydney, NSW, 1997.
[68] Valpar International Corporation, *Manual for MESA 84 - Microcomputer evaluation and screening assessment*, (Revised ed.), Tucson, Arizona: Valpar International Corp, 1984.
[69] Vasudevan, S.V., Role of functional capacity assessment in disability evaluation, *Journal of Back & Musculoskeletal Rehabilitation* **6** (1996), 237–248.
[70] Wesolek, J.S. and McFarlane, F.R., Perceived needs for vocational assessment information as determined by those who utilise assessment results, *Vocational Evaluation & Work Adjustment Bulletin* **24**(2) (1991), 55–60.
[71] Work Evaluation Systems Technology, *WEST 4A user manual: Dynamic upper extremity strength and fatigue tolerance evaluation device*, Long Beach, CA: Work Evaluation Systems Technology, 1989.
[72] Yeomans, S.G. and Liebensen, C., Quantitative functional capacity evaluation: The missing link to outcomes assessment, *Topics in Clinical Chiropractic* **3**(1) (1996), 32–43.

Appendix 1

The following references/sources were those reviewed and analysed for each of the work-related assessments included in this study. While there were more references available for these assessments, only those addressing or commenting on reliability were considered. A more detailed analysis and critique of these references is available from the authors.

Acceptable Maximum Effort (**AME**)

Khalil, T.M., Goldberg, M.L., Asfour, S.S., Moty, E.A., Rosomoff, R.S. and Rosomoff, H.L., Acceptable maximum effort (AME): A psychophysical measure of strength in back pain patients, *Spine* **12**(4) (1987), 372–376.

Applied Rehabilitation Concepts (**ARCON**)

Hasten, D.L., Johnston, F.A. and Lea, R.D., Validity of the Applied Rehabilitation Concepts (ARCON) system for lumbar range of motion, *Spine* **20**(11) (1995), 1279–1283.

AssessAbility

Coupland, M., *AssessAbility manual*, Austin, Texas: IME AssessAbility Inc., 1995.

Rucker, K.S., Wehman, P. and Kregel, J., *Analysis of functional assessment instruments for disability/rehabilitation programs* (Summary report SSA Contract No. 600-95-21914), Richmond, VA: Virginia Commonwealth University, 1996.

Blankenship Functional Capacity Evaluation

Blankenship, K.L., *The Blankenship system functional capacity evaluation: The procedure manual*, (2nd ed.), Macon, GA: The Blankenship Corporation, 1994.

BTE Work Simulator

Anderson, P.A., Chanoski, C.E., Devan, D.L., McMahon, B.L. and Whelan, E.P., Normative study of grip and wrist flexion strength employing a BTE work simulator, *Journal of Hand Surgery* **15A**(3) (1990), 420–425.

Cathey, M.A., Wolfe, F. and Kleinheksel, S.M., Functional ability and work status in patients with fibromyalgia, *Arthritis Care & Research* **1**(2) (1988), 85–98.

Cetinok, E., Coleman, E., Fess, E.E., Dunipace, K.R. and Renfro, R., Reliability of the BTE work simulator dynamic mode [Abstract], *Journal of Hand Therapy* **8**(1) (1995a), 52–53.

Cetinok, E.M., Renfro, R.R. and Coleman, E.F., A pilot study of the reliability of the dynamic mode of one BTE work simulator, *Journal of Hand Therapy* **8**(3) (1995b), 199–205.

Coleman, E.F., Renfro, R.R., Cetinok, E.M., Fess, E.E., Shaar, C.J. and Dunipace, K.R., Reliability of the manual dynamic mode of the Baltimore Therapeutic Equipment Work Simulator, *Journal of Hand Therapy* **9**(3) (1996), 223–237.

Dunipace, K.R., Reliability of the BTE work simulator dynamic mode [Letter to the editor], *Journal of Hand Therapy* **8**(1) (1995), 42–43.

Fess, E.E., Correction: Instrument reliability of the BTE Work Simulator: A preliminary study, *Journal of Hand Therapy* **6**(2) (1993a), 82.

Fess, E.E., Instrument reliability of the BTE work simulator: A preliminary study [Abstract], *Journal of Hand Therapy* **6**(1) (1993b), 59–60.

Kennedy, L.E. and Bhambhani, Y.N., The Baltimore Therapeutic Equipment work simulator: Reliability and validity at three work intensities, *Archives of Physical Medicine & Rehabilitation* **72** (1991), 511–516.

Matheson, L.N. and Ogden-Niemeyer, L., *Work capacity evaluation: Systematic approach to industrial rehabilitation*, (Revised ed.), Anaheim, CA: ERIC, 1987.

McClure, P.W. and Flowers, K.R., Reliability of BTE work measurements [Abstract], *Journal of Orthopaedic & Sports Physical Therapy* **11**(9) (1990), 420.

McClure, P.W. and Flowers, K.R., The reliability of BTE work simulator measurements for selected shoulder and wrist tasks, *Journal of Hand Therapy* **5**(1) (1992), 25–28.

Niemeyer, L.O., Matheson, L.N. and Carlton, R.S., Testing consistency of effort: BTE work simulator, *Industrial Rehabilitation Quarterly* **2**(1) (1989), 5, 12–13, 27–32.

Trossman, P.B. and Li, P.W., The effect of the duration of intertrial rest periods on isometric grip strength performance in young adults, *Occupational Therapy Journal of Research* **9**(6) (1989), 362–378.

Trossman, P.B., Suleski, K.B. and Li, P.-W., Test-retest reliability and day-to-day variability on an isometric grip strength test using the work simulator, *Occupational Therapy Journal of Research* **10**(5) (1990), 266–279.

Cal-FCP (references to EPIC and Spinal Function Sort listed separately)

Matheson, L.N., Mooney, V., Grant, J.E., Leggett, S. and Kenny, K., Standardised evaluation of work capacity, *Journal of Back & Musculoskeletal Rehabilitation* **6** (1996), 249–264.

Dictionary of Occupational Titles - Residual Functional Capacity (**DOT-RFC**)

Fishbain, D.A., Abdel-Moty, E., Cutler, R., Khalil, T.M., Sadek, S., Rosomoff, R.S. and Rosomoff, H.L., Measuring residual functional capacity in chronic low back pain patients based on the Dictionary of Occupational Titles, *Spine* **19**(8) (1994), 872–880.

EPIC Lift Capacity

Alpert, J., Matheson, L., Beam, W. and Mooney, V., The reliability and validity of two new tests of maximum lifting capacity, *Journal of Occupational Rehabilitation* **1**(1) (1991), 13–29.

Matheson, L.N., Relationships among age, body weight, resting heart rate, and performance in a new test of lift capacity, *Journal of Occupational Rehabilitation* **6**(4) (1996), 225–237.

Matheson, L.N., Danner, R., Grant, J. and Mooney, V., Effect of computerised instructions on measurement of lift capacity: Safety, reliability, and validity, *Journal of Occupational Rehabilitation* **3**(2) (1993a), 65–81.

Matheson, L.N., Mooney, V., Grant, J.E., Affleck, M., Hall, H., Melles, T., Lichter, R.L. and McIntosh, G., A test to measure lift capacity of physically impaired adults. Part 1 - Development and reliability testing, *Spine* **20**(19) (1995), 2119–2129.

ERGOS Work Simulator

Matheson, L.N., Danner, R., Grant, J. and Mooney, V., Effect of computerised instructions on measurement of lift capacity: Safety, reliability, and validity, *Journal of Occupational Rehabilitation* **3**(2) (1993a), 65–81.

Isernhagen Functional Capacity Evaluation

Farag, I., *Functional assessment approaches*, Unpublished Master of Safety Science thesis, University of New South Wales, Kensington, NSW, 1995.

Isernhagen Work Systems, *Reliability and validity of the Isernhagen Work systems Functional Capacity Evaluation*, [Publication of limited circulation available from Isernhagen Work Systems, Duluth, Illinois.] 1996.

Smith, R.L., Therapists' ability to identify safe maximum lifting in low back pain patients during functional capacity evaluation, *Journal of Orthopaedic & Sports Physical Therapy* **19**(5) (1994), 277–281.

Key Method Functional Capacity Assessment

Key Functional Assessments, *Key functional assessment procedures manual*, Minneapolis, MN: Key Functional Assessments, 1986.

Key, G.L., Functional capacity assessment, in: *Industrial therapy*, G.L. Key, Ed., St Louis: Mosby, 1995, pp. 220–253.

Lido WorkSET

Hudak, P., Hannah, S., Knapp, M. and Shields, S., Reliability of isometric wrist extension torque using the LIDO WorkSET for late follow-up of postoperative wrist patients, *Journal of Hand Therapy* **10**(4) (1997), 290–296.

Matheson, L.N., Mangesh, G., Segal, J.H., Grant, J.E., Comisso, K. and Westing, S., Validity and reliability of a new device to simulate upper extremity work demands, *Journal of Occupational Rehabilitation* **2**(3) (1992), 109–122.

Shackleton, T.L., Harburn, K.L. and Noh, S., Pilot study of upper-extremity work and power in chronic cumulative trauma disorders, *Occupational Therapy Journal of Research* **17**(1) (1997), 3–24.

MESA/System 2000

Botterbusch, K.F., *Vocational assessment and evaluation systems: A comparison*, Menomonie, Wisconsin: Materials Development Center, Stout Vocational Rehabilitation Institute, University of Wisconsin-Stout, 1987.

Valpar International Corporation, *Manual for MESA 84 - Microcomputer evaluation and screening assessment*, (Revised ed.), Tucson, Arizona: Valpar International Corp, 1984.

Progressive Isoinertial Lifting Evaluation (PILE)

Hazard, R.G., Reeves, V., Fenwick, J.W., Fleming, B.C. and Pope, M.H., Test-retest variation in lifting capacity and indices of subject effort, *Clinical Biomechanics* **8** (1993), 20–24.

Mayer, T., Gatchel, R., Barnes, D., Mayer, H. and Mooney, V., Progressive isoinertial lifting evaluation: Erratum notice, *Spine* **15**(1) (1990), 5.

Mayer, T.G., Barnes, D., Kishino, N.D., Nichols, G., Gatchel, R.J., Mayer, H. and Mooney, V., Progressive isoinertial lifting evaluation I: A standardised protocol and normative database, *Spine* **13**(9) (1988), 993–997.

Polinsky Functional Capacity Assessment

No references located for reliability.

Physical Work Performance Evaluation (PWPE)

Lechner, D.E., Roth, D.L., Jackson, J.R. and Straaton, K., Inter-rater reliability and validity of a newly developed FCE: The physical work performance evaluation [Abstract], *Physical Therapy* **73**(6 Suppl) (1993), S27.

Lechner, D.E., Jackson, J.R., Roth, D.L. and Straaton, K.V., Reliability and validity of a newly developed test of physical work performance, *Journal of Occupational Medicine* **36**(9) (1994), 997–1004.

Smith, E.B., Rasmussen, A.A., Lechner, D.E., Gossman, M.R., Quintana, J.B. and Grubbs, B.L., The effects of lumbosacral support belts and abdominal muscle strength on functional lifting ability in healthy women, *Spine* **21**(3) (1996), 356–366.

Quantitative Functional Capacity Evaluation (QFCE)

Yeomans, S.G. and Liebensen, C., Quantitative functional capacity evaluation: The missing link to outcomes assessment, *Topics in Clinical Chiropractic* **3**(1) (1996), 32–43.

Singer/New Concepts Vocational Evaluation System (Singer VES)

Cohen, C. and Drugo, J., Test-retest reliability of the Singer vocational evaluation system, *Vocational Guidance Quarterly* **24**(3) (1976), 267–270.

Smith Physical Capacity Evaluation (Smith PCE)

Smith, S.L. and Baxter-Petralia, P., *The physical capacities evaluation: Its use in four models of clinical practice*, Baltimore, MD: Chess Publications, 1992.

Spinal Function Sort

Gibson, L. and Strong, J., The reliability and validity of a measure of perceived functional capacity for work in chronic back pain, *Journal of Occupational Rehabilitation* **6**(3) (1996), 159–175.

Matheson, L.N., Matheson, M.L. and Grant, J., Development of a measure of perceived functional ability, *Journal of Occupational Rehabilitation* **3**(1) (1993b), 15–30.

Valpar Component Work Samples (Valpar CWS)

Barrett, T., Browne, D., Lamers, M. and Steding, E., Reliability and validity testing of Valpar 19, Perth, WA.: AAOT, *Proceedings of the 19th National Conference of the Australian Association of Occupational Therapists* **2** (1997), 179–183.

Botterbusch, K.F., *Vocational assessment and evaluation systems: A comparison*, Menomonie, Wisconsin: Materials Development Center, Stout Vocational Rehabilitation Institute, University of Wisconsin-Stout, 1987.

Trevitt, N., *A test-retest reliability study on the Valpar Component Work Sample 4*, Unpublished Honours thesis, School of Occupational Therapy, Faculty of Health Sciences, University of Sydney, NSW, 1997.

Valpar International Corporation, *Manuals for Valpar Component Work Samples* (*1–11*), Tucson, Arizona: Valpar International Corp, 1974.

WEST Standard Evaluation

Hehir, A., *A study of interrater agreement and accuracy of the WEST Standard Evaluation*, Unpublished Honours thesis, School of Occupational Therapy, Faculty of Health Sciences, University of Sydney, NSW, 1995.

Matheson, L.N., Evaluation of lifting and lowering capacity, *Vocational Evaluation & Work Adjustment Bulletin* **19**(3) (1986), 107–111.

Ryan, A., An interrater agreement and accuracy study on the WEST Standard Evaluation [Abstract], *Australian Occupational Therapy Journal* **43**(3/4) (1996), 185.

Tan, H.L., *Investigation of the concurrent validity of an assessment component of the WEST Standard Evaluation for use within Australian population and the accuracy of the WEST 3 Comprehensive Weight System*, Unpublished Honours thesis, School of Occupational Therapy, Faculty of Health Sciences, Curtin University of Technology, Perth, WA, 1995.

Tan, H.L., *Study of the inter-rater, test-retest reliability and content validity of the WEST Standard Evaluation*, Unpublished Masters thesis, School of Occupational Therapy, Faculty of Health Sciences, Curtin University of Technology, Perth, WA, 1996.

Tan, H.L., Barrett, T. and Fowler, B., Study of the inter-rater, test-retest reliability and content validity of the WEST Standard Evaluation, Perth, WA: AAOT, *Proceedings of the 19th National Conference of the Australian Association of Occupational Therapists* **2** (1997), 245–251.

WEST 4/4A

Innes, E., Hargans, K., Turner, R. and Tse, D., Torque strength measurements: An examination of the interchangeability of results in two evaluation devices, *Australian Occupational Therapy Journal* **40**(3) (1993), 103–111.

Matheson, L.N. and Ogden-Niemeyer, L., *Work capacity evaluation: Systematic approach to industrial rehabilitation*, (Revised ed.), Anaheim, CA: ERIC, 1987.

Work Evaluation Systems Technology, *WEST 4A user manual: Dynamic upper extremity strength and fatigue tolerance evaluation device*, Long Beach, CA: Work Evaluation Systems Technology, 1989.

WEST Tool Sort & Loma Linda University Medical Center (LLUMC) Activities Sort

Work Evaluation Systems Technology, *Loma Linda University Medical Center activities sort*, Long Beach, CA: Work Evaluation Systems Technology, 1984a.

Work Evaluation Systems Technology, *WEST tool-sort*, Long Beach, CA: Work Evaluation Systems Technology, 1984b.

WorkAbility Mark III

Shervington, J., *Workplace capability assessment*, Paper presented at the Australia & New Zealand MODAPTS Association International Conference, Melbourne, Vic, 1990, March.

Work Box

Black, M.K., Nelson, C.E., Maurer, P.A. and Bauer, D.F., Test-retest reliability of the Work Box: A work sample with standard instructions, *Work* **3**(4) (1993), 26–34.

Speller, L., Trollinger, J.A., Maurer, P.A., Nelson, C.E. and Bauer, D.E., Comparison of the test-retest reliability of the Work Box using three administrative methods, *American Journal of Occupational Therapy* **51**(7) (1997), 516–522.

WorkHab Australia Functional Capacity Evaluation

Bradbury, S. and Roberts, D., *WorkHab Australia Functional Capacity Evaluation workshop manual*, Bundaberg, Qld: WorkHab Australia, 1996.