Validity 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 21 December 1998
Accepted 29 January 1999

Insufficient evidence of the validity of work-related assessments is frequently reported as a major concern in occupational rehabilitation. Despite this concern, and the continuing development of new and old assessments, no comprehensive evaluation of the evidence has been conducted. *Objectives:* The purpose of this study was to first determine the extent and quality of available evidence for the validity of work-related assessments, and then where sufficient evidence was available, determine the level of validity. *Study Design:* This study examined available literature and sources in order to review the extent to which validity has been established for 28 work-related assessments. *Results:* The levels of evidence and validity are presented for each assessment. Most work-related assessments have limited evidence of validity. Of those that had adequate evidence, validity ranged from poor to good. There was no instrument that demonstrated moderate to good validity in all areas. Very few work-related assessments were able to demonstrate adequate validity in more than one area, or with more than one study, even when contributory evidence was included. *Conclusion:* With this study clinicians will be able to examine their options with regard to the validity of the assessments they choose to use.

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

1. Introduction

The ongoing concern for clinicians in occupational rehabilitation is the usefulness of assessment results in guiding the safe and swift return to work of injured workers. To be useful to clinicians the results must provide valid and reliable information to enable appropriate clinical decisions. For more than a decade concerns over the usefulness of current work-related assessments have fuelled the call for work-related assessments to demonstrate valid and reliable results [1,31,40,41,44, 52,61,62,78,110,111].

While there is limited evidence for either validity or reliability, it appears that the validity of work-related assessments has been examined even less than reliability [44]. Ten commercial work-related assessment systems used in the United States of America (USA) were reviewed for evidence of validity and reliability by Lechner et al. [62]. They found that only three assessments had content validity (Isernhagen FCE, Smith PCE and Valpar Component Work Samples), only one had criterion-related validity (Smith PCE), and only one had construct validity for one component of the assessment (Sweat). These results should be of major concern to clinicians. Unfortunately, the sources of information on which some of the results were based were not reported, and so it is not possible to review the original studies. There were also a number of published studies not used.

Since the review by Lechner et al. [62] further assessments have been developed, and existing systems revised, modified and updated. Some assessments are no longer commercially available although they may remain in use. King et al. [60] conducted a more recent review, also of ten commercial work-related assessments, only three of which were included in the Lechner et al. [62] study (Blankenship FCE, Isernhagen FCE, and Key FCA). The remaining seven assessments had either been developed since 1991 or were not included by Lechner et al. King et al. [60] reported that only two assessments (ErgoScience PWPE, and WEST-EPIC/Cal-FCP lift capacity section) had inter-rater and intra-rater reliability studies published, and only one had a published validity study (ErgoScience PWPE). There was, however, no discussion or critique of either of these studies, or those conducted on the ERGOS. A further three assessments indicated published research associated with them (ARCON, Isernhagen FCE, and WorkAbility Mark III), however, there was no reference to the sources of these publications.

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

Work 13 (1999) 125–152
ISSN 1051-9815 / $8.00

Both Lechner et al. [62] and King et al. [60] reviewed a wide range of aspects associated with a limited range of work-related assessments, including evidence of reliability and validity. Neither of these reviews, however, focussed exclusively on these issues and so was unable to explore reliability or validity in depth. For this reason, the current study has examined the existing available literature in detail in order to review the extent to which validity has been established for a wide range of work-related assessments. A previous paper examined the evidence of reliability for the same range of assessments [45].

1.1. Validity

Validity is usually considered to be the extent to which an instrument measures what it is intended to measure [83]. The validity of a test refers to the appropriateness, meaningfulness and usefulness of the specific inferences made from the test results [22]. Validity refers to the results of a test and how they are interpreted, not the instrument itself. Successfully determining an injured worker's ability to safely return to work performing specified suitable duties is based on a valid interpretation of test results.

Validity is inferred from research findings and applied experience, using personal, as well as generally accepted standards [22]. Work-related assessments are rarely totally invalid or valid; rather their validity is a matter of degree that can best be considered as good, moderate, poor or unknown.

A confusing and inappropriate use of the term *validity* occurs in some work-related assessments. The terms *validity profile* (e.g., Blankenship FCE), *valid, conditionally valid, conditionally invalid* and *invalid* effort (e.g., Key FCA) are used by some systems. These terms do not refer to the validity of the instrument or test battery results, but rather the level of effort exerted by the client performing the assessment. They are used to describe the level or sincerity of effort exerted by a client and are not related to the measurement concept of validity. Clinicians should be aware of this use of the term and note that "there is no peer-reviewed scientific justification for the use of the term *validity profile* as that term relates to functional testing" [34, p. 351].

Validity depends on the purpose of the assessment, and therefore the test objectives. It is not a universal characteristic of an assessment [83]. Rather, it is always specific to some particular use [33]. Further, no single measure is sufficient from which to determine an assessment's validity. These aspects imply that multiple studies of the various forms of validity are required and that validity must be evaluated within the context of the test's intended purpose and a specific population [83]. Ideally, clinicians should be able to determine the circumstances of their need to use a work-related assessment then select an assessment that has demonstrated validity for a similar specific population when used for a similar defined purpose and within a similar specific context.

1.2. Types of validity

There are several forms of validity that may be determined. These are face, content, criterion-related (concurrent and predictive) and construct validity. All of these forms of validity are relevant to work-related assessments.

1.2.1. Face validity

Face validity is evident when a work-related assessment appears to measure what it intends to measure and it is considered a plausible method to do so [83]. It is about "logical inferences – what appears to be sensible logical reasoning" [16, p. 30]. For example, a work sample such as the Valpar Component Work Sample 16 (Drafting), which requires the client to copy from drawings using equipment associated with drafting, may be considered to have high face validity because of the clear association with the perceived job requirements of drafting.

Face validity can be established by a panel or group of experts who examine the assessment and reach a consensus that it does or does not represent a particular concept [20]. However, face validity can also be established by clients, therapists and consumers of test results such as insurers, managers and employers.

Face validity needs to be evaluated for a particular purpose. For example, a work-related assessment's face validity may be considered in terms of its ability to adequately assess the duties, tasks, task elements or elemental motions required for a particular job [43]. The ability to adequately assess the duties of a car mechanic is a very different concept to that of assessing the specific task elements or skills required to perform these duties. A work-related assessment may be considered to have poor face validity to simulate the duties of a car mechanic, but good face validity to assess the task elements of lifting, reaching and using hand-tools required to perform the duties of a car mechanic. The concepts on which the determination of good or poor face validity is made are clearly different.

Face validity is considered by some authors to be part of content validity [20,22], while some do not consider it as a form of validity at all [33]. It is the most basic and least rigorous form of validity and has no standard for determining whether an instrument has sufficient or adequate face validity [22,83]. As a result, it is not sufficient to only have evidence of face validity, as it is considered to be subjective and scientifically weak [83]. While relying on face validity as the only form of validity can be criticised as being insufficient for a work-related assessment, not establishing this form of validity can also be a problem. Without obvious face validity clients, therapists and consumers of test results may consider an assessment irrelevant and unacceptable [83].

1.2.2. Content validity

Content validity is the degree to which test items represent the performance domain the test is intended to measure. For example, one work-related assessment may include items examining whole body physical demands such as lifting, carrying, climbing and walking, while another focuses on hand and upper limb coordination and dexterity.

Content validity is usually determined by a panel of experts who examine the relationship between test objectives and test items, or by knowledge of the normal practices used [53,106]. Content validity is not usually indicated by a statistical measure, but rather inferred from expert judgements, and certain logical procedures [22]. It considers whether the test incorporates a representative sample of the components of the task in question, such as a work-related assessment incorporating relevant job demands [60].

To determine content validity it is necessary to establish the rationale for the test, provide operational definitions of the test variables and identify the specific objectives of the instrument [83]. The assessment can then be examined at both the specific item and more general test level [106]. At the more detailed level each item is examined to determine the extent to which it is a measure of the content domain, while at a broader level the entire range of test items can be considered in terms of its representativeness of the content domain [106]. Content validity is considered to be a prerequisite for criterion-related and construct validity and should generally be established before either of these [106].

The need to clearly define an assessment's rationale and objectives is extremely important in the area of work-related assessments. The level of a work-related assessment (i.e., role, activity, task, skill or body system) [43] and the stated objectives will influence the determination of content validity. For example, an instrument comprising of tests of various tasks (e.g., lifting, carrying, climbing, etc) may be considered to have poor face or content validity if the objective is to determine an individual's ability to return to the job of a hairdresser. However, if the objective is to determine an individual's ability to perform a variety of work-related physical tasks, then the validity may be much higher.

It may appear that face and content validity are similar, and indeed face validity has been described as a component of content validity [83]. Some have tried to differentiate face and content validity based on the time of the validity determination. For example, Portney and Watkins [83] suggest face validity is determined after an assessment is developed, while content validity is established as part of the planning and development process of the instrument. However, a more useful method of differentiating may be to view face validity as demonstrating the general relevance of an instrument to the overall purpose of the assessment. This logical relationship is clear to all users of the instrument and consumers of the results. Content validity is the detailed relationship between the specific parts or subtests of an instrument and the components of the tasks or activities in question. It is of more concern to specialists using the instrument, rather than lay consumers of the results. It is the examination of both the general and specific aspects of an instrument that are considered by clinicians when selecting an assessment [53].

1.2.3. Criterion-related validity

Criterion-related validity is the systematic demonstration of the extent to which test performance is related to some other valued measure of performance or external criterion [22,33]. It is comprised of concurrent and predictive validity and is considered to be "the most practical approach to validity testing and the most objective. It is based on the ability of one test to predict results obtained on another test" [83, p. 73]. Scores from the work-related assessment being evaluated (i.e., the target test) are compared and correlated with those from the criterion measure. Concurrent and predictive validity are described as follows:

– *Concurrent validity* examines the correlation between two or more measures given to the same subjects at approximately the same time so that both reflect the same incident of behaviour [83]. The new measure is compared to an existing, valued measure or 'gold standard'. This approach is

useful when the target test is new or untested and is being proposed as an alternative assessment to the criterion measure because of ease of administration, efficiency, practicality and/or safety [83].

– *Predictive validity* compares a subject's performance at the initial time of testing to performance obtained at a future date with another highly valued measure or 'gold standard' [22]. Establishing predictive validity is essential for clinical decision-making and would indicate that the target test was a valid predictor of a future criterion score [83]. For work-related assessments a client's success when returning to work is a highly valued criterion. While many assessments claim an ability to do so, very few have demonstrated this predictive validity.

It is assumed that the criterion measure selected is an established and valid indicator of the variable of interest [83]. In order to establish the utility of the criterion measure it should generally demonstrate acceptable test-retest reliability, have relevance to the behaviour being measured in the target test and be independent of the target test's results [83]. The valued criterion of return-to-work is certainly a valid indicator that is relevant and independent of test results. It may be argued that return-to-work does not have demonstrated test-retest reliability as a criterion measure, however, it is certainly considered a 'gold standard' by which the results of the target test are compared.

Selecting an appropriate criterion measure can be a difficult task, especially if the constructs are abstract or if there is no recognised 'gold standard' [83]. A common problem encountered with work-related assessments is that many have non-existent, or at best limited evidence of reliability or validity. This makes selection of an acceptable criterion measure particularly problematic. As a result, new assessments are compared with pre-existing instruments that do not have adequate evidence of reliability or validity, or with other new instruments that are assumed to measure similar constructs, but for which there is also no adequate evidence for reliability and validity.

1.2.4. Construct validity

Construct validity is the extent to which a test can be shown to measure a hypothetical construct [22]. For example, a work-related assessment may be considered to have some support for construct validity if it is able to differentiate between clients who are able to lift safely and those who do not, where the construct being measured is safe lifting ability.

There is no single method to determine construct validity, but rather an accumulation of evidence, often over numerous studies [81,83]. Methods used in collecting evidence for construct validity include the following:

– *Known Groups Method* is the most general type of evidence and involves the ability of the test results to discriminate between groups which are known to be different (e.g., different diagnostic groups; different age groups; different occupational groups) in a theoretically appropriate manner [22,33,83]. For example, the Valpar Component Work Sample (VCWS) 6 (Independent Problem-Solving) was able to differentiate between subjects with and without brain damage [8], providing support for construct validity.

– *Correlation with other tests* involves the examination of the degree of convergence and/or divergence with other tests that are presumed to measure the same or different constructs or traits [22, 33,83]. It is also referred to as a multitrait-multimethod matrix [83]. It may appear that convergent and divergent validity are similar to concurrent validity in that all compare the target test with other instruments. The purposes, however, differ. The focus of convergent/divergent validity is on the construct examined rather than the comparison of results with a criterion measure or gold standard. Concurrent validity assumes that the tests are examining the same construct.

  * Convergent validity compares the target test with other measures believed to reflect the same construct(s) [83]. If the same construct is reflected in both tests the results should correlate highly. The MESA Interest Survey, for example, has good convergent validity when compared with the USES Interest Inventory [50], with both instruments examining the construct of occupational interest.
  * Discriminant or divergent validity compares the target test with other measures believed to assess different characteristics or traits [83]. A low correlation is expected in this case. For example, an assessment of lifting and carrying ability would not be expected to correlate highly with one examining clerical skill.

– *Hypothesis testing* involves identifying specific hypotheses that support the theoretical basis of the test and the constructs included [83]. For example, it may be hypothesised that following a work

hardening program a client will improve performance on a number of measures. By comparing scores pre- and post-treatment the test results should change (or remain stable) under the various treatment/intervention conditions in an hypothesised manner [33]. There are numerous examples of change following treatment programs as measured by various work-related assessments (e.g., [57,73,75,79,87]).

– *Factor analysis* is an approach that examines the factor structure of a test by testing different populations to ensure that the internal structure of the test is not different between diagnostic subgroups (i.e., the factors or constructs are stable in different situations) [22]. Each factor represents a group of test items or behaviours related to each other but not to other factors within the test, and reflects a different theoretical component of the overall construct [83]. For example, there may be a number of test items related to hand and upper limb function and considered to be a factor. This factor should be unrelated to test items focussed on standing and walking, which would be considered part of a different factor.

Construct validity is the broadest type of validity, and content and criterion-related validity may be used to support construct validity [22,83]. It is necessary to define the content domain that represents the construct and to also define the constructs according to a theoretical context [83]. Demonstrating good construct validity enables greater generalisation over various populations and situations [55].

1.2.5. Screening

Using assessments for screening purposes enables early detection of disease or dysfunction [83]. A cutoff point is usually established from which the presence or absence of a target condition is established [83]. Screening may be considered as part of construct validity because its aim is to differentiate between groups by determining whether a person does or does not have a particular condition.

The validity of screening assessments is determined by examining the test's sensitivity and specificity to a target condition. *Sensitivity* is the test's ability to obtain a true positive result, that is a positive result when the condition is actually present. *Specificity* is the test's ability to obtain a true negative result, which is a negative result when the condition is absent [83]. Positive and negative predictive values can also be calculated. These values provide an estimate of the likelihood that a person who tests positive actually has the condition (positive predictive value) or the converse, a person tests negative and does not exhibit the condition (negative predictive value) [83].

There is often a trade-off between sensitivity and specificity. As the criterion or cutoff point for determining the presence of a specific condition becomes less stringent, there will be greater sensitivity, but less specificity. The reverse also applies where a more stringent cutoff point will give less sensitivity and greater specificity [83]. The clinical decision that is required is what levels of sensitivity and specificity are acceptable. Consideration needs to be given to the consequences of obtaining false positives (identifying the presence of a condition when it is absent) and false negatives (not identifying the condition when it is actually present) [83].

This concept has important implications for work-related assessments that incorporate tests to determine a client's level or sincerity of effort. It may be preferable to have very stringent criteria that reduce the sensitivity (reduce the incidence of true positives) but increase the specificity (increase the incidence of true negatives) to avoid inappropriate labelling of individuals as producing a sub-maximal effort. This may result in an increase in false negatives, where a sub-maximal effort is considered maximal, however, this may be preferable to incorrectly identifying a maximal effort as sub-maximal (false positive).

While there are a number of tests used to determine the level or sincerity of effort, there are few that have specified cutoff points and only one study was identified that examined the sensitivity and specificity of these criteria [51].

1.3. 'Good' validity

Unlike reliability, validity is not as straightforward to establish, due to difficulty verifying measurement inferences [83]. "For many variables there are no obvious rules or formulas for judging that a test is indeed measuring the critical property of interest" [83, p. 71].

As indicated previously statistical measures or standards are not usually used to establish face validity [22, 83]. Some qualitative interpretation can, however, be made, indicating whether good, moderate or poor face validity exists (Table 1). Content validity is established by expert opinion, but some statistical techniques have been used to support that opinion. Thorn and Deitz [106] suggest the use of an index of item-

objective congruence. This measure is a procedure for the analysis of judgements of content experts and was originally introduced by Rovinelli and Hambleton (1977, cited in [106]). The index allows examination of content validity at the test-item level and has a range from -1.00 to $+1.00$, indicating the worst possible to perfect congruence between the test-item and relevant test-objective or domain [106]. A score of $\geqslant +0.70$ is considered acceptable for item inclusion, while items with indices between $+0.50$ and $+0.69$ should be examined individually to decide to accept, revise or reject items [106].

Others methods, such as percentage agreement and the kappa (κ) coefficient, have also been suggested, however, there are limitations with these approaches [106]. Percentage agreement can give spuriously high results because it does not account for chance agreement, while κ requires many judgements to be made by the content experts [106]. None of these quantitative methods, including the index of item-objective congruence, have been used to determine the content validity of work-related assessments.

The level of content validity can be considered in the same way as face validity, with good, moderate and poor levels according to agreement by content experts reviewing the specific items in relation to the relevant test objectives (Table 1).

For criterion-related validity (concurrent and predictive) similar statistics as for content validity are used (i.e., percentage agreement, correlation and kappa coefficients). McFadyen and Pratt [78] indicate that the interpretation of correlation coefficients is similar to that used for reliability. However, although Portney and Watkins [83] suggest guidelines for the interpretation of reliability coefficients, they do not indicate that these guidelines are appropriate for validity. Some studies have used percentage agreement to examine the predictive validity of work-related assessments (e.g., [66]), however, there is the concern that this form of analysis does not account for chance. It has been suggested that 70% agreement is required for clinical utility and 90% agreement is considered good, however, this was with reference to inter-rater agreement, rather than validity of results [39].

Convergent and discriminant validity of work-related assessments also use correlation coefficients to analyse data. The correlation coefficients are incorporated into a multitrait-multimethod matrix (e.g., [48, 50,108]). Convergent validity should have correlations that are moderately high, but not too high, and statistically significant (Anastasi, 1988, cited in [50]). If there is high correlation between a new test and an already available test, without additional advantages such as speed or ease of administration, then the new test may unnecessarily duplicate an existing instrument. Correlations for construct validity of 0.60 or greater are considered "high", while those between 0.30 and 0.60 are "moderate to good" [91] (Table 1). Discriminant validity is only examined if there is sufficient evidence of convergent validity [48].

Other aspects of construct validity, such as discrimination between known groups and demonstrating change following treatment, use a wide variety of statistical procedures to analyse data. Inferential statistics such as t-tests, Wilcoxon and analysis of variance are used to determine whether group or treatment differences exist (Table 1). It is beyond the scope of this paper to examine all the inferential statistics used. Clinicians should, however, be aware of the assumptions and appropriate use of these procedures to determine if valid conclusions are drawn.

Sensitivity and specificity are calculated as the proportion or percentage of subjects who either actually do (sensitivity) or do not (specificity) have the condition being tested [83]. Similarly predictive value is also calculated as a percentage. There are, however, no guidelines regarding acceptable levels of sensitivity, specificity or predictive value. It is dependent on the need to identify the existence of a particular condition. The clinician must therefore determine the importance of identifying the condition and the subsequent sensitivity and specificity required in an instrument.

1.4. Validity of work-related assessments

All forms of validity are appropriate for work-related assessments. Face and content validity are required to demonstrate the relevance of the assessment to the client, therapist, employer, insurer and others involved in assisting injured workers to return to work safely and quickly. Criterion-related validity is important to demonstrate that the results of a work-related assessment can predict successful return to work (predictive validity), as well as being at least as efficient as existing techniques in determining work ability (concurrent validity). Construct validity provides evidence that work-related assessments can discriminate between different groups, such as those with and without back pain, detect change in injured workers following treatment, and adequately assess the constructs on which the instrument is based. If a work-related assessment is also to be used for screening purposes, such as whether a client

Table 1
Levels of validity

Type of validity	Level of validity	Interpretation of level
Face Validity	Unknown	Insufficient evidence upon which to base a sound judgement.
	Poor	Most experts, clients &/or test result users consider there is little relation between the test and what it is intended to measure.
	Moderate	Most experts, clients &/or test result users consider there is some relationship between the test and what it is intended to measure, however, some relevant components are not included.
	Good	Most experts, clients &/or test result users agree that the test measures what is intended, and all relevant components are included.
Content Validity	Unknown	Insufficient evidence upon which to base a sound judgement.
	Poor	Most experts consider there is little relation between the test and what it is intended to measure.
	Moderate	Most experts consider there is some relationship between the test and what it is intended to measure, however, some relevant components are not included.
	Good	Most experts agree that the test measures what is intended, and all relevant components are included.
Criterion Validity	Unknown	Insufficient evidence upon which to base a sound judgement.
	Poor	Statistical evidence suggests there is little similarity between the test and criterion measure (e.g., percentage agreement $< 70\%, \kappa \leqslant 0.40, r \leqslant 0.50$).
	Moderate	Statistical evidence suggests there is some similarity between the test and criterion measure (e.g., percentage agreement $\geqslant 70\%, \kappa > 0.40, r > 0.50$).
	Good	Statistical evidence suggests there is substantial similarity between the test and criterion measure (e.g., percentage agreement $\geqslant 90\%, \kappa > 0.60, r > 0.75$).
Construct Validity	Unknown	Insufficient evidence upon which to base a sound judgement.
	Poor	Statistical evidence suggests a poor ability to differentiate between groups or interventions (small effect size), or poor convergence between similar tests (e.g., $r < 0.30$), or poor divergence between similar tests.
	Moderate	Statistical evidence suggests a moderate ability to differentiate between groups or interventions (medium effect size), or moderate convergence between similar tests (e.g., $r \geqslant 0.30$), or moderate divergence between similar tests.
	Good	Statistical evidence suggests a good ability to differentiate between groups or interventions (large effect size), or good convergence between similar tests (e.g., $r \geqslant 0.60$), or good divergence between similar tests.

is exerting maximum effort, then acceptable sensitivity and specificity must also be demonstrated.

In a review of functional assessment literature and methods conducted for the USA Social Security Administration (SSA) instruments were automatically excluded from further review if there was no evidence of validity or reliability, and no citations of research [89]. This demonstrates the need for acceptable and accessible evidence of validity for work-related assessments. Unfortunately, as with evidence for reliability, there is a dearth of studies examining various aspects of validity for the work-related assessments currently in use and commercially available.

2. Method

This study utilised the same methodology as that used to determine the extent of evidence for reliability of work-related assessments and examined the same instruments [45]. The following sources of information were accessed:

- 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.

Twenty-eight work-related assessments were included in this study from a total of 55 that were identified. The selection criteria for inclusion in the study were work-related assessments 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 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 [107], and the Physio-Tek and Sweat FCA, both referred to by Lechner et al. [62]. 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 validity of work-related assessments

Each work-related assessment included in this study was examined for evidence from validity studies, as well as evidence from other studies that contributed validity evidence (contributory evidence). 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, analy-

ses and interpretation of results to enable a judgement to be made on the acceptability of the validity of the assessment studied. As validity requires an accumulation of evidence, often over multiple studies of the various forms of validity, studies that did not specifically examine validity, but used a work-related assessment as one of a range of instruments within a study were also examined. The inclusion of studies that do not specifically examine the validity of work-related assessments is appropriate because of the lack of specific validity studies for many of the instruments included in this study. All evidence that can contribute to establishing the validity of an instrument should be considered. This includes studies where the focus, for example, may be on determining the efficacy of a work hardening program. Using a work-related assessment to determine change in subjects over the course of the program contributes to the instrument's construct validity by demonstrating its ability to detect change. By examining these studies, as well as those that specifically examine validity, it is possible to build a more detailed picture of the overall validity of work-related assessments. Appendix 1 identifies each of the sources used.

The levels of evidence for the validity of work-related assessments included in this review were categorised into six broad categories using the same definitions as were used for reliability [45] (Table 2). The lowest level (Level 0) indicates that no evidence for validity was identified. Level 1 indicates that the developers of the assessment relied on previous studies conducted on various aspects of the assessment. The assumption made by the test developers is that the previous studies demonstrated acceptable validity and so justifies the inclusion of the particular aspect of the test. Generalising acceptable validity for some aspects to all components of the assessment is dangerous. Furthermore there may have been no critical review of the previous studies before accepting the results reported.

Level 2 indicates that although there may be some report of validity, 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. Examples of Level 3 evidence are often, but not always, abstracts of conference presentations where limited space precludes greater detail being provided. Sufficient detail for the evaluation of results consists of a description of the type of validity 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 validity from a number of these levels. It should be noted, however, that although there may be an adequate level of evidence (i.e., the validity of an assessment has been examined and reported in adequate detail in a peer-reviewed forum), this does not indicate that the level of validity is acceptable for clinical purposes.

For each work-related assessment included in this study all available evidence of validity was located and examined, including contributory evidence. 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). The level of validity was then determined as good, moderate, poor or unknown based on the interpretation of measures of validity described previously (see Tables 1 and 4).

3. Results

A summary of the level of evidence for validity that could be located for the range 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 validity is reported in Table 4.

3.1. Studies with insufficient evidence for validity (levels 0–3)

No formal validity studies were identified (Level 0) for the AME, Cal-FCP, Key FCA, Lido WorkSET, PILE, Polinsky FCA, WEST 4/4A, WEST Tool Sort and LLUMC Activity Sort, or WorkHab (Australia). All assessments except WorkHab, however, had evidence contributing to validity in some form. This contributory evidence ranged between Levels 2 to 5 and was usually for face/content and/or construct validity.

Evidence for AssessAbility, which is based on MTM (Methods-Time-Measurement) data, was considered to be at Level 1 as it assumes that MTM data has "content, context and predictive validity" [18, p. 5-1] based on previous research. While the use of predetermined time-motion standards such as MTM may be an appropriate basis from which to develop an assessment, no

Table 2
Levels of evidence for validity

Level	Description
0	No validity demonstrated or reported.
1	Validity is assumed from previous studies conducted on aspects now incorporated into the current assessment. Previous studies may be in either a non-peer-reviewed or peer-reviewed forum.
2	Validity is reported, but there is no detail provided to enable examination of the results. May be in either a non-peer-reviewed or peer-reviewed forum.
3	Validity 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. Often, but not always, an abstract of a conference presentation.
4	Validity 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	Validity, with sufficient detail to enable examination of the results, is reported and published in a peer-reviewed forum (i.e., peer-reviewed journal).

formal validity studies or other contributory evidence have been reported on AssessAbility.

The Isernhagen FCE has content validity reported with respect to the US Department of Labor's physical demands [47,60], however no detail is provided for further examination (Level 2). Contribution to content validity [62] and construct validity [24,46,47] is provided at Levels 2 to 4. Percentages of clients who had returned to work following an Isernhagen FCE were reported [46,47], however, no other statistical analysis of the data was undertaken making it impossible to determine the predictive validity of the assessment (Level 3). The study by Farag [24] attempted to compare psychophysical and kinesiophysical lifting capacity in injured and uninjured subjects (Level 4). Psychophysical results were significantly higher than the kinesiophysical results for both the injured and uninjured groups. Unfortunately, there was no comparison of the lifting capacity of injured subjects with uninjured subjects by Farag. Subsequent analysis by one of the authors of the current study (EI) found no significant difference between the lifting capacity of injured and uninjured subjects for either psychophysical or kinesiophysical approaches. This suggests that there is moderate construct validity in differentiating between techniques for determining a safe lifting end-point (i.e., kinesiophysical versus psychophysical), however, there is no support for the ability to differentiate between injured and uninjured subjects.

The Blankenship FCE and the QFCE both have evidence of content validity at Level 3. The QFCE has no other evidence of validity. The Blankenship FCE, however, has contributory evidence for content and construct validity at Levels 2, 3 and 5. Examination of maximal and submaximal effort in clients was the purpose of studies associated with construct validity [10, 54]. From a database of over 6,000 subjects, Blankenship [10] reported the percentage of clients not exerting good effort as determined by the assessment's 'validity profile'. No other analyses were undertaken, and as this was a conference abstract (Level 3) it is not possible to examine the results in any further detail. Kaplan et al. [54] only used a small number of sub-tests from the Blankenship FCE to determine maximal and submaximal effort. As the results of the physical demand subtests were not reported, it is not possible to compare results from subjects deemed to be exerting a maximal or submaximal effort. Therefore, it is not possible to comment on any aspects of validity.

The AME assessment has no formal studies of its validity, however, a study examining pre- and post-treatment change in lifting capacity of clients with low back pain provides contributory evidence supporting good construct validity (i.e., ability to determine effect of treatment) [57]. There are no other studies, however, which support this finding.

The Cal-FCP has some contributory evidence (Level 5) supporting its criterion-related and construct validity, based on the inclusion of the Spinal Function Sort and EPIC Lift Capacity as components of the overall assessment [74].

Contributory evidence for the Lido WorkSET provides support for good construct validity in its ability to differentiate between healthy subjects and those with chronic upper extremity cumulative trauma disorder [94]. There is also some contributory evidence for isotonic strength as a predictor of work capacity, although the authors of the study do not feel that this is the case for isometric strength [26,114].

Table 3 Continued

Assessment	Types of validity			
	Face/Content	Criterion-related	Construct	Screening
			5 (*#6 - correlations with neuropsychological tests; compared workers & subjects with mental illness*) *5* (*#4 - compared hand injured & healthy groups*) *5* (*#8, #9 - neck pain; sick/not sick listed*) *5* (*#5 - change in earning capacity with RA*) *5* (*#1 - functional loss attributable to hand impairment*)	
WEST Std Eval.	**4** *2* (*DOT physical demands*) *3* (*physical demands*)	**4** (MHRWS & 3-D motion analysis) *5* (*prediction for RTW*) *5* (*WEST & Lido trunk dynamometer & future work injury*)	**0** *4* (*norms for different occupational groups, injury types, F/M*) *4* (*compared US & Aust. "norms" - considered to be concurrent V*) *5* (*pre/post treatment change - LBP*) *5* (*pre/post treatment change - body mechanics instruction*) *5* (*pre/post treatment change - LBP*)	**0**
WEST 4/4A	**0**	**0** *5* (*WEST 4 & BTE*) *5* (*WEST 4A & FAST*)	**0**	**0**
WEST Tool & LLUMC Activity Sorts	**0** *3* (*Chinese translation of Tool Sort*)	**0**	**0**	**0**
WorkAbility Mk 3	**2** (DOT physical demands) **3, 4** (MODAPTS Activity Groups)	**0**	**0**	**0**
Work Box	**0**	**0**	**0** *5* (*differences between F/M & job-related experience*)	**0**
WorkHab	**0**	**0**	**0**	**0**

N.B. Numbers (**0–5**) in **bold** type indicate level of evidence for validity, while numbers (*0–5*) in *italic* type indicate a contribution to validity. 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.

The Polinsky FCA has contributory evidence suggesting poor predictive validity when clients with low back injuries attempt to predict their actual lifting capacity and standing tolerance [82]. The authors of the study suggest that this finding supports the use of work-related assessments to assist in promoting a safe return to work. This conclusion, however, is based on the assumption that the assessment is able to determine a safe level of work, which was not the focus of the study.

The Work Box has contributory evidence supporting moderate construct validity for its ability to discriminate between level of experience with tasks requiring manual dexterity, and also between genders [101]. There are no other studies, however, which support this finding.

The WEST 4/4A has some contributory evidence for concurrent validity indicating that while there is fair correlation with the BTE and moderate correlation with the FAST, the BTE is not significantly different from the WEST 4/4A [113], while the FAST is different from the WEST 4/4A [42]. It should be noted that Wolf et al. [113] have misinterpreted a low shared variance between the WEST 4/4A and the BTE as indicating a demonstration of significant difference between the two instruments, despite reporting no significant difference. This contributory evidence indicates poor concurrent

Table 4 Continued

Assessment	Types of validity			
	Face/Content	Criterion-related	Construct	Screening
			Moderate (pre/post treatment change - LBP) *Poor - moderate* (pre/post treatment change - body mechanics instruction) *Good* (pre/post treatment change - LBP)	
WEST 4/4A	Unknown	*Poor* (WEST 4 & BTE) *Poor* (WEST 4A & FAST)	Unknown	Unknown
WEST Tool & LLUMC Activity Sorts	Unknown	Unknown	Unknown	Unknown
WorkAbility Mk 3	**Moderate-good** (MODAPTS Activity Groups)	Unknown	Unknown	Unknown
Work Box	Unknown	Unknown	*Moderate* (differences between F/M & job-related experience)	Unknown
WorkHab	Unknown	Unknown	Unknown	Unknown

N.B. The assessments in **bold** are those with evidence of validity at Level 4 or 5, while those in *italic* type indicate a contribution to validity 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.

validity between the WEST 4/4A and both the BTE and the FAST.

The PILE has no formal validity studies, however, there is extensive published literature that provides contributory evidence for its construct validity. Based on several studies there is evidence of good construct validity for the PILE's ability to detect change in lifting capacity following various types of work hardening and functional restoration programs [19,38,75,85]. There is poor correlation between the PILE and Cybex Liftask, indicating that the tests measure different aspects of lifting and cannot be substituted for each other [75, 77].

3.2. Studies with sufficient evidence for validity (levels 4–5)

The ARCON, BTE Work Simulator, DOT-RFC, EPIC Lift Capacity, ERGOS Work Simulator, MESA/ System 2000, PWPE, Singer/New Concepts VES, Smith PCE, Spinal Function Sort, Valpar Component Work Samples, WEST Standard Evaluation and WorkAbility Mk III all have evidence of validity at Levels 4 and 5. Some assessments also have evidence at lower levels (i.e., Levels 1 to 3). All assessments, with the exceptions of the Singer VES and WorkAbility Mk III, have contributory evidence of validity. This is most often for construct validity, with published studies examining pre- and post-treatment change, or differences between various groups of subjects.

The ARCON was examined for criterion-related validity by comparing lumbar range of motion results with the 'gold standard' of dual inclinometry in a group of healthy subjects [36]. Correlations between the two assessments were highly variable, ranging from poor to good for various sub-tests [36]. The authors concluded that the validity criterion in the American Medical Association Guides to the Evaluation of Permanent Impairment was "not met for active or passive SLR for either sex on the ARCON" [36, p. 1282]. This conclusion of poor concurrent validity was supported in a later study [37]. Static lift, push and pull components of the ARCON were found to be significantly improved in a group of subjects with low back dysfunction who were tested before and after a six week work hardening program [87]. This finding contributes to good construct validity of these components of the ARCON.

The BTE Work Simulator is one of the most extensively researched work-related assessments with respect to criterion-related validity, and has many studies that contribute to establishing its construct validity (Level 5). Interestingly, it appears that face and content validity are assumed, rather than being formally evaluated, with only a cursory overview (Level 2) of the physical demands covered by the BTE [62].

Several BTE attachments set up to simulate various levels and types of work have been compared to the

actual work demands to establish moderate criterion-related validity (Level 5 – [56,112]. Both studies found that the BTE tended to underestimate the energy requirements (VO_2 and heart rate) of the work tasks. Poor concurrent validity was found when the BTE was compared with arm cranking for VO_2 and heart rate [7] prompting the authors to recommend that as many actual work simulation tasks as possible should be included in a test battery to ensure a comprehensive assessment. The BTE attachment #162 has been compared with the Jamar dynamometer when determining grip strength in a number of studies and found to have good concurrent validity [5,35,58].

There are no studies that formally investigate the BTE's construct validity, however, there are numerous studies (Level 5) that contribute to establishing it. The contributory evidence indicates moderate construct validity for discrimination between obviously different groups (different methods of upper extremity exercise [9]; e.g., comparison of patient and healthy groups [4]), however, this does not appear to be the case when there is greater similarity between groups ([14]; e.g., comparison of groups with fibromyalgia and rheumatoid arthritis [14,27]; comparison of surgical approaches [32]). Using an impairment rating as a criterion for predicting functional loss as determined by a number of measures, including the BTE has also been found to have poor predictive validity [88].

Studies contributing to the use of the BTE as a screening tool for determining the level of effort exerted by clients [59,80] should be interpreted with caution. Both studies examined the coefficients of variation (CVs) produced by a number of BTE attachments and suggested cutoff points to differentiate between maximal and sub-maximal effort. Neither study determined the predictive values, sensitivity or specificity of the suggested cutoff points. It should also be noted that the use of CVs for determining sincerity of effort is actively discouraged by Lechner et al. [63], who state that the use of CVs for this purpose is unsubstantiated in the literature.

The DOT-RFC is the only assessment that has content validity established at Level 5 [25]. This is in relation to the DOT physical demands. In the same study a factor analysis found the physical demands assessed fell into four major groups (mobility/strength, pushing/pulling, tolerance and manual dexterity) accounting for 62.4% of the variance in results. The authors concluded that this supported the design of the test battery, providing some evidence of construct validity.

The EPIC Lift Capacity test and its precursor the Progressive Lift Capacity (PLC) test have been used in several studies of concurrent (criterion-related) validity [2,68,70]. In all these studies, however, the EPIC or PLC was considered the criterion test against which other assessments were compared. This appears to stem from the assumption that a functional dynamic lift, as performed in the EPIC, has greater face validity than isokinetic lifts or movements performed on the Lido Lift and Lido Passive Back Machine, and isometric lifts performed on the ERGOS Work Simulator. Using the EPIC as the criterion measure or 'gold standard' against which to compare other assessments does not appear justified when the validity of the EPIC has not been established, despite its good to excellent reliability [2, 72]. While the EPIC, an isoinertial lifting assessment, has moderate correlation with isokinetic and isometric lifts, it is not possible to comment on the concurrent validity of this assessment.

Good construct validity was established for the EPIC's ability to measure change in lifting ability of younger and middle-aged back injured subjects following treatment [73]. The ability to predict lifting capacity based on subject age, body weight, height, and resting heart rate also supports construct validity [69]. The EPIC was unable, however, to determine any difference in lifting capacity based on the use of a lumbar support belt [86].

There are encouraging results regarding the EPIC's "indicators of sincere effort" to differentiate between maximal and submaximal effort, with the authors reporting excellent positive (94.44%) and good negative (80.00%) predictive values [51]. This study, however, is reported as a conference abstract and so has only limited information available for examination (Level 3).

The ERGOS Work Simulator has been examined for criterion-related validity [23,70]. In one study human instructions were found to have better correlation with static lift performance that computerised instructions [70]. The same study found there were higher correlations between the ERGOS static lifts and a test of dynamic lifting (EPIC) at knuckle level, but not at elbow level with computerised instructions. Human instructions, however, produced high correlations between the two instruments at either knuckle or elbow level. When the ERGOS was compared with other established tests (therapist physical evaluation, workshop tasks and Valpar Component Work Samples), it was found that there was wide variation in the correlation and κ coefficients computed [23]. There was substantial agreement ($\kappa = 0.66$) between the ERGOS results and the final physical activity rating compiled

by the vocational evaluator, which was interpreted as demonstrating the concurrent validity of the ERGOS in comparison to current methods of evaluation [23].

While construct validity has not been specifically examined for the ERGOS, two studies contribute to this area [17,97]. Neither study, however, supported the constructs examined. Cooke et al. [17, p. 761] considered there was no "useful predictive value when applied to an individual" in the sub-tests examined because the range in performance in normal subjects is so wide. The variability of individual performance also precluded the use of CVs to determine subject effort in a study by Simonsen [97].

MESA/System 2000 has the most extensive study of its construct validity of any of the assessments reviewed. Convergent and divergent validity were examined for MESA and a range of vocational assessments, interest checklists and intelligence tests [48–50,102]. Overall, there was support for the construct validity of MESA's academic achievement, general educational development, interest survey and aptitude scores. Most correlations were moderate ($r = 0.40$ to 0.60), however, Stoelting [102] considered that the aptitude scores fell short of offering predictive validity. A study examining clients' perceptions of MESA also contributed to the face validity of the assessment [11].

The PWPE has been examined for some aspects of concurrent validity, with moderate correlation between the overall work level recommended and the level of work currently performed [64,65]. A Level 3 study reported an 87% agreement between PWPE results and actual work status 3 and 6 months post-discharge [66] providing some support for criterion-related validity. Other reported studies (Level 3 and 5) contribute to the construct validity of the PWPE when examining the differences in coordination tasks and lifting produced by different age groups, males and females and varying anthropometric measures [6,12,84,98].

The Singer/New Concepts VES demonstrates moderate criterion-related validity with 82% of job samples having correlations (r_s) at or above 0.50 when compared with employment success in jobs specifically in the occupational groups associated with the job sample [28]. This was confirmed in a later study [29].

The Smith PCE is considered to be a valid predictor of return to work (RTW) status (criterion-related validity – Level 5) [99,100]. This conclusion was based, however, on comparison between assessment results and a client completed questionnaire identifying if the client had returned to work or not. Smith et al. [100] acknowledge that there was a high non-RTW rate (73%) and also a high non-return rate of the questionnaire (42% returned) which may have affected results and limits their generalisability.

The Spinal Function Sort demonstrates good convergence (construct validity) with a number of pain, self-efficacy and work scales [30]. Further support for the instrument's construct validity is provided by studies that demonstrate an ability to differentiate between subjects with acute, sub-acute and chronic low back pain [71,103].

The Valpar Component Work Samples have a wide range of studies (Levels 2 to 5) examining all aspects of validity. Only the VCWS 19 has had face/content validity studied in detail (Level 4 – [3]). It was rated as having poor face validity because the expert panel did not consider that all of the critical job demands of a stores/shipping clerk were covered by the work sample. The VCWS 8 was also compared to an actual job (mail officer) and found to lack critical job demands (Level 3 – [93]). These critical job demands were at a task rather than skills level. If the physical demands or skills, rather than the job tasks were examined there may be a different outcome [43].

Convergent (construct) validity was examined between a number of Valpar work samples and the General Aptitude Test Battery (GATB) aptitude scores (Level 5 – [91]). Examining the pattern of intercorrelations, it was concluded that there was support for the construct validity of VCWS 4, tentative support for VCWS 8 and 9 and no clear support for VCWS 6. The other work samples (VCWS 7, 10 and 11) "seem to be measuring other areas of general behaviour than that measured by the GATB subtests" [91, p. 23]. VCWS 6 was able to differentiate between subjects with and without brain damage with 78.9% accuracy [8].

VCWS 4 and 8 are also able to differentiate between groups of subjects, providing support for construct validity. There was a significant difference ($p < 0.05$) between subjects with hand injuries and matched controls when assessed using the VCWS 4 [15], however, there was no suggestion of scores that may be considered to discriminate between the two groups. Schult et al. [92] attempted to determine if there was a difference between subjects who were sick-listed and those who were not when assessed by VCWS 8 and 9. They reported no logical pattern between successfully completing the work samples and not being sick-listed. A subsequent analysis by one of the authors of the current study (EI), however, found subjects who were not sick-listed performed significantly better on the VCWS 8 (i.e., completed it successfully) than those who were

sick-listed ($x = 11.58, df = 1, p < 0.001$). It was not possible, however, to calculate this for VCWS 9.

Moderate criterion-related (concurrent) validity was demonstrated between VCWS and both therapists' evaluation and workshop tasks [23]. There was poor predictive validity, however, when an impairment rating was used to predict functional loss as determined by a range of measures including the VCWS 1 [88].

The WEST Standard Evaluation has poor content validity based on expert opinion [104,105]. Experts consider that the assessment does not provide adequate information on a person's lifting and lowering capacity.

The WEST Standard Evaluation has poor to fair concurrent validity when the Measurement of High Risk Work Style is compared with the criterion measure of three-dimensional motion analysis [39,90]. This may be due to 3-D motion analysis being much more sensitive to slight changes in movement than the naked eye. There is support, however, of moderate construct validity related to the ability to detect change following intervention [13,76,79].

As with the EPIC, the WEST Standard Evaluation has been used as the criterion measure with which to compare the results of isokinetic trunk testing [21]. This selection again appears to be based on the face validity of the instrument, rather than other forms of established validity and without good reliability demonstrated. It is therefore not possible to adequately evaluate the results of the study.

WorkAbility Mk III has moderate to good content validity [95,96]. The study is considered by its authors, however, to be evidence of concurrent validity. Given that the study compared employers' analyses of various jobs with the MODAPTS-based 'activity groups' used in WorkAbility Mk 3, it would appear that the study was examining the content, rather than concurrent validity of the assessment.

4. Discussion

4.1. Level of validity

Face and content validity appear to be rarely formally established for the majority of work-related assessments. It would seem that most consider a work-related assessment to demonstrate adequate content validity when it is possible to identify most, if not all physical demands as described in the Dictionary of Occupational Titles within the instrument [60,62]. This determination is usually made at the most cursory level without support or justification for the acceptance of these criteria. It also assumes that inclusion of job task elements at the skill level, such as lifting, standing and climbing, will be adequate for determining an individual's ability to perform the duties and tasks associated with a specific job [43].

Only the DOT-RFC and WorkAbility Mk III demonstrate moderate to good content validity. The VCWS 19 and WEST Standard Evaluation have also had content validity established through expert panels, however, it was found to be poor for both assessments. This is in contrast to both King et al. [60] and Lechner et al. [62] who report "good" content validity for the ERGOS, Isernhagen FCE, Key FCA, PWPE, Valpar CWS and WorkAbility Mk III, but without justification for these decisions, other than comparison with the Dictionary of Occupational Titles' physical demands.

While determination of content validity has been commonly based on expert opinion, it may assist developers and users of work-related assessments to consider more structured methods such as determining item-objective congruence when establishing content validity in the future. Given the importance of demonstrating face and content validity to users and consumers of work-related assessments, further formal research in this area is warranted.

Criterion-related validity was the most common formally evaluated type of validity examined in work-related assessments. There was moderate validity demonstrated for the ErgoScience PWPE, Singer/New Concepts VES and Smith PCE when compared with the ability to return to work. While this was at a very general level for the Smith PCE (i.e., return to work, no return to work), the PWPE considered the specific return to work level (i.e., sedentary, light, medium, heavy, very heavy) and the Singer/New Concepts VES identified the specific job type.

When compared with work simulation or workshop tasks the BTE and ERGOS Work Simulators, and the Valpar Component Work Samples demonstrated moderate concurrent validity. It was recommended, however, that as many work simulation tasks as possible be included in a test battery to ensure a comprehensive assessment [7]. This would, however, clearly depend on the purpose of the assessment. Where the specific job requirements are known, it would not be necessary to assess a wide range of simulated work tasks, although it may be necessary if no specific job has been identified.

Work-related assessments, such as the ARCON and WEST Standard Evaluation, had poor criterion-related validity when compared with instruments used to mea-

sure specific aspects of movement, such as the dual inclinometer and three-dimensional motion analysis system. The poor outcomes may be the result of either an incompatible criterion being selected for comparison, or the criterion being too sensitive. Good criterion-related validity was only demonstrated when a work-related assessment was compared with a similar instrument (e.g., BTE #162 compared with the Jamar dynamometer – [5]). This highlights the difficulty of attempting to establish the validity of work-related assessments. It also indicates the need to carefully select an appropriate and acceptable criterion standard.

Construct validity was rarely formally evaluated. However, approximately half of the work-related assessments included in this study had some contributory evidence of construct validity. This was most commonly in the form of demonstrating a treatment effect or differentiating between different groups. The PILE, for example, has demonstrated an ability to detect change in lifting ability following treatment in a number of studies [19,38,75,85], supporting its construct validity for this purpose. The BTE appears to be able to detect differences between different groups at a gross level (e.g., between healthy subjects and those with fibromyalgia – [14]), but not when the differences are more subtle (between two surgical approaches for brachial plexus lesions – [4]; e.g., between subjects with fibromyalgia and those with rheumatoid arthritis – [14]).

Convergent and divergent aspects of construct validity were only addressed for MESA/System 2000, Spinal Function Sort and Valpar Component Work Samples. MESA/System 2000 and VCWS are both based on the same system used to analyse jobs in the Dictionary of Occupational Titles [109], and can therefore be compared with other instruments using the same constructs. The Spinal Function Sort was compared with other measures of similar constructs. Given that many work-related assessments are reportedly based on the physical demands of the DOT, it would seem reasonable that these constructs could be examined.

While a number of work-related assessments purport to identify subjects producing maximal or sub-maximal performance, no Level 4 or 5 studies examining this feature were located for any work-related assessment. There is some promising research, however, which begins to address this concern [51]. Only VCWS 6 (Independent Problem-Solving) has demonstrated an ability to screen subjects for cognitive deficits [8].

4.2. Limitations of the study

It is recognised that a limitation of this study is that evidence of validity 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 similar difficulty in locating contributory evidence is also acknowledged. When the focus of a study is determining the efficacy of treatment, for example, there is no clear or obvious indication that a particular work-related assessment is used to measure outcome. Therefore, despite these studies being published, it is possible that some may not have been identified and included in this current study.

It is also recognised that work-related assessments such as AssessAbility, Cal-FCP and WorkHab are relatively recent additions to the range of work-related assessments (published in 1995, 1994 and 1996, respectively) and so there has been limited time in which to conduct studies examining the reliability and validity of these assessments.

Return-to-work systems and legislation associated with occupational rehabilitation and workers' compensation vary within and between countries where work-related assessments are used. This will influence the reason for conducting a work-related assessment, how the results are reported and used, and the selection of assessments to meet identified needs. These factors will influence the type of validity studies undertaken as well as the generalisability of results to different contexts.

4.3. Validity and reliability

As highlighted in a previous paper [45], reliability and validity are independent continua that may be positively or negatively associated. This association will depend on the context of the assessment, the level of the assessment (i.e., role/job, activity/duty, task or skill/task element) and the type of validity considered.

For example, a work-related assessment that focuses on the skill or task element level, such as the EPIC Lift Capacity test, can determine test-retest and inter-rater reliability relatively easily. A good level of reliability can be expected, and has in fact been established for this work-related assessment [72]. Evidence for face,

content, criterion-related and construct validity may also be relatively straightforward to establish because variables can be controlled, and test components studied in detail. Reliability and validity for this type of assessment is therefore positively correlated (i.e., good validity is associated with good reliability).

Workplace-based assessments, however, focus on the role level of performance. Both test-retest and inter-rater reliability are much more difficult to determine in this situation due to the non-standardised and variable nature of the assessment, and the difficulty in replicating the test environment and other extraneous variables. The performance of the actual job in the real work environment results in justifiably high face and content validity, although criterion-related and construct validity may be more difficult to establish. Face and content validity of work-related assessments at the role level are, therefore, negatively correlated with reliability (i.e., good face and content validity may be associated with poor reliability).

Demonstration of acceptable reliability is usually considered a precursor to demonstrating an instrument's validity [83], that is reliability and validity are positively associated. For work-related assessments, however, this may not always be the case. The level of the assessment (i.e., role, activity, task or skill), the context of the assessment and the type of validity examined can influence the correlation between reliability and validity.

There may be a tendency for clinicians to modify and adapt work-related assessments when the purpose of the assessment is inconsistent with the level of the instrument. Clinicians modify and adapt standardised assessments when the instrument does not meet their requirements [67]. For example, when an instrument assesses performance at a task or skill level, but the referral question requires an answer with respect to role or activity performance, poor face or content validity may be identified. In an attempt to improve the face and content validity of an instrument, clinicians may add or remove components of the assessment, include simulations of necessary tasks and activities, or go to the workplace. This area has not been examined and requires extensive further research.

5. Conclusion

As with reliability, most work-related assessments have limited evidence of validity. A number had insufficient evidence on which to base an assessment of the level of validity. Of those that had adequate evidence, validity ranged from poor to good. Work-related assessments with adequate evidence of moderate to good validity included some attachments of the BTE Work Simulator, DOT-RFC, EPIC Lift Capacity, ERGOS Work Simulator, MESA/System 2000, PWPE, Singer/New Concepts VES, Smith PCE, Spinal Function Sort, Valpar CWS and WorkAbility Mk III. Other instruments had contributory evidence that began to establish moderate to good validity. These included AME, ARCON, Cal-FCP, Isernhagen FCE, Lido Work-SET, PILE, WEST Standard Evaluation and the Work Box.

There was, however, no instrument that demonstrated moderate to good validity in all areas. Very few work-related assessments were able to demonstrate adequate validity in more than one area, or with more than one study, even when contributory evidence was included. This highlights the need for further research to be conducted in this area. Test developers, clinicians and academics are strongly encouraged to continue investigating the validity of work-related assessments.

The acceptance of work-related assessments on the basis of their longevity in the marketplace and clinic should not be assumed to equate with adequate validity. With this review clinicians are now able to examine their options with regard to the validity of the work-related assessments they choose to use.

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] 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.

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

[4] Beaton, D.E., Dumont, A., Mackay, M.B. and Richards, R.R., Steindler and pectoralis major flexorplasty: A comparative analysis, *Journal of Hand Surgery* **20A**(5) (1995a), 747–756.

[5] Beaton, D.E., O'Driscoll, S.W. and Richards, R., Grip strength testing using the BTE work simulator and the Jamar dynamometer: A comparative study, *Journal of Hand Surgery* **20A**(2) (1995b), 293–298.

[6] Bevington, J., Warner, L., Hyde, S.D.A., Lechner, D.E. and Gossman, M.R., Performance values on four coordination tasks for healthy, working-aged adults [Abstract], *Physical Therapy* **74**(5) (1994), S99.

[7] Bhambhani, Y., Esmail, S. and Britnell, S., The Baltimore Therapeutic Equipment work simulator: Biomechanical and physiological norms for three attachments in healthy men, *American Journal of Occupational Therapy* **48**(1) (1994), 19–25.

[8] Bielecki, R.A. and Growick, B., Validation of the Valpar independent problem-solving work sample as a screening tool for brain damage, *Vocational Evaluation & Work Adjustment Bulletin* **17**(2) (1984), 59–61.

[9] Blackmore, S.M., Beaulieu, D., Baxter-Petralia, P. and Bruening, L., A comparison study of three methods to determine exercise resistance and duration for the BTE work simulator, *Journal of Hand Therapy* **1**(4) (1988), 165–171.

[10] Blankenship, K.L., The Blankenship FCE system behavioural profile: A four year retrospective study [Abstract], Brisbane, Qld: Australian Physiotherapy Association, *Proceedings of the 1996 National Physiotherapy Congress of the Australian Physiotherapy Association* (1996), 111–112.

[11] Bordieri, J.E. and Musgrave, J., Client perceptions of the Microcomputer Evaluation and Screening Assessment, *Rehabilitation Counseling Bulletin* **32**(4) (1989), 342–345.

[12] Buckley, E., Rasmussen, A.A., Lechner, D., Gossman, M.R., Quintana, J.B. and Grubbs, B., The effects of lumbosacral support belts and abdominal muscle strength on functional lifting ability in healthy women [Abstract], *Physical Therapy* **74**(5) (1994), S27.

[13] Carlton, R.S., The effects of body mechanics instruction on work performance, *American Journal of Occupational Therapy* **41**(1) (1987), 16–20.

[14] 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.

[15] Cederlund, R., The use of dexterity tests in hand rehabilitation, *Scandinavian Journal of Occupational Therapy* **2**(3–4) (1995), 99–104.

[16] Clemson, L. and Fitzgerald, M.H., Understanding assessment concepts within the occupational therapy context, *Occupational Therapy International* **5**(1) (1998), 18–34.

[17] Cooke, C., Dusik, L.A., Menard, M.R., Fairburn, S.M. and Beach, G.N., Relationship of performance on the ERGOS work simulator to illness behaviour in a workers' compensation population with low back versus limb injury, *Journal of Occupational Medicine* **36**(7) (1994), 757–762.

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

[19] Curtis, L., Mayer, T.G. and Gatchel, R.J., Physical progress and residual impairment quantification after functional restoration. Part III: Isokinetic and isoinertial lifting capacity, *Spine* **19**(4) (1994), 401–405.

[20] Dane, F.C., *Research methods*, Pacific Grove, CA: Brooks/Cole Publishing, 1990.

[21] Dueker, J.A., Ritchie, S.M., Knox, T.J. and Rose, S.J., Isokinetic trunk testing and employment, *Journal of Occupational Medicine* **36**(1) (1994), 42–48.

[22] Dunn, W., Reliability and validity, in: *Developing norm-referenced standardised tests*, L.J. Miller, Ed., New York: Haworth Press, 1989.

[23] Dusik, L.A., Menard, M.R., Cooke, C., Fairburn, S.M. and Beach, G.N., Concurrent validity of the ERGOS work simulator versus conventional functional capacity evaluation techniques in a workers' compensation population, *Journal of Occupational Medicine* **35**(8) (1993), 759–767.

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

[25] 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.

[26] Ford, D., Kwak, A. and Wolfe, L.D., Grip strength decrease and recovery following isotonic exercise [Abstract], *Journal of Hand Therapy* **3**(1) (1990), 36.

[27] Fraulin, F.O., Louie, G., Zorrilla, L. and Tilley, W., Functional evaluation of the shoulder following latissimus dorsi muscle transfer, *Annals of Plastic Surgery* **35**(4) (1995), 349–355.

[28] Gannaway, T.W. and Sink, J.M., The relationship between the vocational evaluation system by Singer and employment success in occupational groups, *Vocational Evaluation & Work Adjustment Bulletin* **11**(2) (1978), 38–45.

[29] Gannaway, T.W., Sink, J.M. and Becket, W.C., A predictive validity study of a job sample program with handicapped and disadvantaged individuals, *Vocational Guidance Quarterly* **29**(1) (1980), 4–11.

[30] 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.

[31] Gibson, L. and Strong, J., A review of functional capacity evaluation practice, *Work* **9**(1) (1997), 3–11.

[32] Goldner, R.D., Howson, M.P., Nunley, J.A., Fitch, R.D., Belding, N.R. and Urbaniak, J.R., One hundred eleven thumb amputations: Replantation vs revision, *Microsurgery* **11**(3) (1990), 243–250.

[33] Gronlund, N.E., *Measurement and evaluation in teaching*, (4th ed.), New York: Macmillan, 1981.

[34] 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.

[35] Harvey, P. and Gench, B., A comparison of static grip strength measurements taken on the Jamar dynamometer and the BTE [Abstract], *Journal of Hand Therapy* **6**(1) (1993), 53–54.

[36] 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.

[37] Hasten, D.L., Lea, R.D. and Johnston, F.A., Lumbar range of motion in male heavy laborers on the Applied Rehabilitation Concepts (ARCON) system, *Spine* **21**(19) (1996), 2230–2234.

[38] Hazard, R.G., Fenwick, J.W., Kalisch, S.M., Redmond, J., Reeves, V., Reid, S. and Frymoyer, J.W., Functional restoration with behavioural support: A one-year prospective study of patients with chronic low-back pain, *Spine* **14**(2) (1989), 157–161.

[39] 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, The University of Sydney, Sydney, NSW, 1995.

[40] Innes, E., *Work evaluation systems - What are our current options?* Paper presented at the 6th State Conference of the NSWAOT, Mudgee, NSW, October 1993.

[41] 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.

[42] 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.
[43] Innes, E. and Straker, L., A clinician's guide to work-related assessments: 2 - Design problems, *Work* **11**(2) (1998a), 191–206.
[44] Innes, E. and Straker, L., A clinician's guide to work-related assessments: 3 - Administration and interpretation problems, *Work* **11**(2) (1998b), 207–219.
[45] Innes, E. and Straker, L., Reliability of work-related assessments, *Work* **13** (1999), 107–124.
[46] Isernhagen, S.J., Contemporary issues in functional capacity evaluation, in: *The comprehensive guide to work injury management*, S.J. Isernhagen, Ed., Gaithersburg, MD: Aspen, 1995, pp. 410–429.
[47] Isernhagen Work Systems, *Reliability and validity of the Isernhagen Work Systems Functional Capacity Evaluation*, Duluth, Ill: Isernhagen Work Systems, 1996.
[48] Janikowski, T.P., Berven, N.L. and Bordieri, J.E., Validity of the Microcomputer Evaluation Screening and Assessment aptitude scores, *Rehabilitation Counseling Bulletin* **35**(1) (1991), 38–51.
[49] Janikowski, T.P., Bordieri, J.E. and Musgrave, J.R., Construct validation of the academic achievement and general educational development subtests of the Microcomputer Evaluation Screening and Assessment (MESA), *Vocational Evaluation & Work Adjustment Bulletin* **23**(1) (1990a), 11–16.
[50] Janikowski, T.P., Bordieri, J.E., Shelton, D. and Musgrave, J., Convergent and discriminant validity of the Microcomputer Evaluation Screening and Assessment (MESA) interest survey, *Rehabilitation Counseling Bulletin* **34**(2) (1990b), 139–149.
[51] Jay, M.A., Lamb, J.M., Watson, R.L. and Young, I.A., Sensitivity and specificity of the indicators of sincere effort of the EPIC Lift Capacity test on a previously injured population [Abstract], *Physical Therapy* **78**(5) (1998), S64.
[52] 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.
[53] Johnston, M.V., Keith, R.A. and Hinderer, S.R., Measurement standards for interdisciplinary medical rehabilitation, *Archives of Physical Medicine & Rehabilitation* **73**(12-S) (1992), 3–23.
[54] Kaplan, G.M., Wurtele, S.K. and Gillis, D., Maximal effort during functional capacity evaluations: An examination of psychological factors, *Archives of Physical Medicine & Rehabilitation* **77**(2) (1996), 161–164.
[55] Keith, R.A., Functional assessment measures in medical rehabilitation: Current status, *Archives of Physical Medicine & Rehabilitation* **65** (1984), 74–78.
[56] 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.
[57] 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.
[58] King, J.W. and Berryhill, B.H., A comparison of two static grip testing methods and its clinical applications: A preliminary study, *Journal of Hand Therapy* **1** (1988), 204–208.
[59] King, J.W. and Berryhill, B.H., Assessing maximum effort in upper extremity functional testing, *Work* **1**(3) (1991), 65–76.
[60] King, P.M., Tuckwell, N. and Barrett, T.E., A critical review of functional capacity evaluations, *Physical Therapy* **78**(8) (1998), 852–866.
[61] Krefting, L.M. and Bremner, A., Work evaluation: Choosing a commercial system, *Canadian Journal of Occupational Therapy* **52**(1) (1985), 20–24.
[62] Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.
[63] Lechner, D.E., Bradbury, S.F. and Bradley, L.A., Detecting sincerity of effort: A summary of methods and approaches, *Physical Therapy* **78**(8) (1998), 867–888.
[64] 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.
[65] Lechner, D.E., 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) (1993), S27.
[66] Lechner, D.E., Sheffield, G.L., Page, J.J. and Jackson, J.R., Predictive validity of a functional capacity evaluation: The physical work performance evaluation [Abstract], *Physical Therapy* **76**(5) (1996), S81.
[67] Managh, M.F. and Cook, J.V., The use of standardised assessment in occupational therapy: The BaFPE-R as an example, *American Journal of Occupational Therapy* **47**(10) (1993), 877–884.
[68] Matheson, L., Mooney, V., Caiozzo, V., Jarvis, G., Pottinger, J., DeBerry, C., Backlund, K., Klein, K. and Antoni, J., Effect of instructions on isokinetic trunk strength testing variability, reliability, absolute value, and predictive validity, *Spine* **17**(8) (1992), 914–921.
[69] 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.
[70] 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.
[71] 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.
[72] 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) (1995a), 2119–2129.
[73] Matheson, L.N., Mooney, V., Holmes, D., Leggett, S., Grant, J.E., Negri, S. and Holmes, B., A test to measure lift capacity of physically impaired adults. Part 2 - Reactivity in a patient sample, *Spine* **20**(19) (1995b), 2130–2134.
[74] 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.
[75] Mayer, T.G., Barnes, D., Nichols, G., Kishino, N.D., Coval, K., Piel, B., Hoshino, D. and Gatchel, R.J., Progressive isoinertial lifting evaluation II: A comparison with isokinetic lifting in a disabled chronic low-back pain industrial population, *Spine* **13**(9) (1988), 998–1002.
[76] Mayer, T.G., Gatchel, R.J., Kishino, N., Keeley, J., Capra, P., Mayer, H., Barnett, J. and Mooney, V., Objective assessment

of spine function following industrial injury: A prospective study with comparison group and one-year follow-up, *Spine* **10**(6) (1985), 482–493.

[77] Mayer, T.G., Mooney, V., Gatchel, R.J., Barnes, D., Terry, A., Smith, S. and Mayer, H., Quantifying postoperative deficits of physical function following spinal surgery, *Clinical Orthopaedics & Related Research* **244** (1989), 147–157.

[78] McFadyen, A.K. and Pratt, J., Understanding the statistical concepts of measures of work performance, *British Journal of Occupational Therapy* **60**(6) (1997), 279–284.

[79] Moran, M. and Strong, J., Outcomes of a rehabilitation programme for patients with chronic back pain, *British Journal of Occupational Therapy* **58**(10) (1995), 435–438.

[80] 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.

[81] Ottenbacher, K.J., Methodological issues in measurement of functional status and rehabilitation outcomes, in: *Functional assessment and outcome measures for the rehabilitation health professional,* S.S. Dittmar and G.E. Gresham, Eds., Gaithersburg, Maryland: Aspen, 1997, pp. 17–26.

[82] Piela, C.R., Hallenberg, K.K., Geoghegan, A.E., Monsein, M.R. and Lindgren, B.R., Prediction of functional capacities, *Work* **6**(2) (1996), 107–113.

[83] Portney, L.G. and Watkins, M.P., *Foundations of clinical research: Applications to practice,* Norwalk, Connecticut: Appleton & Lange, 1993.

[84] Prim, J.F., Shealy, S.A., Lechner, D.E., Gossman, M.R. and Bradley, E., Factors influencing the lifting ability of healthy females 20 to 35 years of age [Abstract], *Physical Therapy* **73**(6) (1993), S51.

[85] Rainville, J., Ahern, D.K., Phalen, L., Childs, L.A. and Sutherland, R., The association of pain with physical activities in chronic low back pain, *Spine* **17**(9) (1992), 1060–1064.

[86] Reyna, J.R., Leggett, S.H., Kenney, K., Holmes, B. and Mooney, V., The effect of lumbar belts on isolated lumbar muscle: Strength and dynamic capacity, *Spine* **20**(1) (1995), 68–73.

[87] Robert, J.J., Blide, R.W., McWhorter, K. and Coursey, C., The effects of a work hardening program on cardiovascular fitness and muscular strength, *Spine* **20**(10) (1995), 1187–1193.

[88] Rondinelli, R.D., Dunn, W., Hassanein, K.M., Keesling, C.A., Meredith, S.C., Schulz, T.L. and Lawrence, N.J., A simulation of hand impairments: Effects on upper extremity function and implications toward medical impairment rating and disability determination, *Archives of Physical Medicine & Rehabilitation* **78**(12) (1997), 1358–1363.

[89] 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.

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

[91] Saxon, J.P., Spitznagel, R.J. and Shellhorn-Schutt, P.K., Intercorrelations of selected VALPAR Component Work Samples and General Aptitude Test Battery scores, *Vocational Evaluation & Work Adjustment Bulletin* **16**(1) (1983), 20–23.

[92] Schult, M., Söderback, I. and Jacobs, K., Swedish use and validation of Valpar work samples for patients with musculoskeletal neck and shoulder pain, *Work* **5**(3) (1995), 223–233.

[93] Sen, S., Fraser, K., Evans, O.M. and Stuckey, R., A comparison of the physical demands of a specific job and those measured by standard functional capacity assessment tools, in: *Ergonomics and human environments: Proceedings of the 27th Annual Conference of the Ergonomics Society of Australia,* V. Propovic and M. Walker, Eds., Coolum, Qld: Ergonomics Society of Australia, 1991, pp. 263–268.

[94] 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.

[95] Shervington, J. and Balla, J., Screening workplace capabilities for competitive employment: Report on workplace feedback, in: *Industrial engineering in occupational health: ANZMA seminars,* (Vol. 3, No. 1), J.M. Farrell, Ed., Melbourne, Vic.: Australia & New Zealand MODAPTS Association, 1994, pp. 31–65.

[96] Shervington, J. and Balla, J., WorkAbility Mark III: Functional assessment of workplace capabilities, *Work* **7**(3) (1996), 191–202.

[97] Simonsen, J.C., Coefficient of variation as a measure of subject effort, *Archives of Physical Medicine & Rehabilitation* **76**(6) (1995), 516–520.

[98] 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.

[99] Smith, S.L., Cunningham, S. and Weinberg, R., Predicting reemployment of the physically disabled worker, *Occupational Therapy Journal of Research* **3**(3) (1983), 178–179.

[100] Smith, S.L., Cunningham, S. and Weinberg, R., The predictive validity of the functional capacities evaluation, *American Journal of Occupational Therapy* **40**(8) (1986), 564–567.

[101] 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.

[102] Stoelting, C., A study of the construct validity of the MESA, *Vocational Evaluation & Work Adjustment Bulletin* **23**(3) (1990), 85–91.

[103] Sufka, A., Hauger, B., Trenary, M., Bishop, B., Hagen, A., Lozon, R. and Martens, B., Centralization of low back pain and perceived functional outcome, *Journal of Orthopaedic & Sports Physical Therapy* **27**(3) (1998), 205–212.

[104] 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.

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

[106] Thorn, D.W. and Deitz, J.C., Examining content validity through the use of content experts, *Occupational Therapy Journal of Research* **9**(6) (1989), 334–346.

[107] Tramposh, A.K., The functional capacity evaluation: Measuring maximal work abilities, *Occupational Medicine: State of the Art Reviews* **7**(1) (1992), 113–124.

[108] Tryjankowski, E.M., Convergent-discriminant validity of the Jewish Employment and Vocational Service system, *Journal of Learning Disabilities* **20**(7) (1987), 433–435.
[109] U.S. Department of Labor, Employment & Training, *The revised handbook for analyzing jobs*, Indianapolis, IN: JIST Works, 1991.
[110] Vasudevan, S.V., Role of functional capacity assessment in disability evaluation, *Journal of Back & Musculoskeletal Rehabilitation* **6** (1996), 237–248.
[111] 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.
[112] Wilke, N.A., Sheldahl, L.M., Dougherty, S.M., Levandoski, S.G. and Tristani, F.E., Baltimore Therapeutic Equipment Work Simulator: Energy expenditure of work activities in cardiac patients, *Archives of Physical Medicine & Rehabilitation* **74**(4) (1993), 419–424.
[113] Wolf, L.D., Klein, L. and Cauldwell-Klein, E., Comparison of torque strength measurements on two evaluation devices, *Journal of Hand Therapy* **2** (1987), 24–27.
[114] Wolf, L.D., Matheson, L.N., Ford, D.D. and Kwak, A.L., Relationships among grip strength, work capacity and recovery, *Journal of Occupational Rehabilitation* **6**(1) (1996), 57–70.

Appendix 1

The following references/sources were those reviewed and analysed for each of the work-related assessments included in the study. While there were more references available for these assessments, only those addressing or commenting on validity were considered.

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.

Hasten, D.L., Lea, R.D. and Johnston, F.A., Lumbar range of motion in male heavy laborers on the Applied Rehabilitation Concepts (ARCON) system, *Spine* **21**(19) (1996), 2230–2234.

Robert, J.J., Blide, R.W., McWhorter, K. and Coursey, C., The effects of a work hardening program on cardiovascular fitness and muscular strength, *Spine* **20**(10) (1995), 1187–1193.

AssessAbility

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

Blankenship Functional Capacity Evaluation

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

Blankenship, K.L., The Blankenship FCE system behavioural profile: A four year retrospective study, Brisbane, Qld: A.P.A., *Proceedings of the 1996 National Physiotherapy Congress of the Australian Physiotherapy Association* (1996), 111–112.

Kaplan, G.M., Wurtele, S.K. and Gillis, D., Maximal effort during functional capacity evaluations: An examination of psychological factors, *Archives of Physical Medicine & Rehabilitation* **77**(2) (1996), 161–164.

Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.

BTE Work Simulator

Beaton, D.E., Dumont, A., Mackay, M.B. and Richards, R.R., Steindler and pectoralis major flexorplasty: A comparative analysis, *Journal of Hand Surgery* **20A**(5) (1995), 747–756.

Beaton, D.E., O'Driscoll, S.W. and Richards, R., Grip strength testing using the BTE work simulator and the Jamar dynamometer: A comparative study, *Journal of Hand Surgery* **20A**(2) (1995), 293–298.

Bhambhani, Y., Esmail, S. and Britnell, S., The Baltimore Therapeutic Equipment work simulator: Biomechanical and physiological norms for three attachments in healthy men, *American Journal of Occupational Therapy* **48**(1) (1994), 19–25.

Blackmore, S.M., Beaulieu, D., Baxter-Petralia, P. and Bruening, L., A comparison study of three methods to determine exercise resistance and duration for the BTE work simulator, *Journal of Hand Therapy* **1**(4) (1988), 165–171.

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.

Esmail, S., Bhambhani, Y. and Britnell, S., Gender differences in work performance on the Baltimore Therapeutic Equipment work simulator, *American Journal of Occupational Therapy* **49**(5) (1995), 405–411.

Fraulin, F.O., Louie, G., Zorrilla, L. and Tilley, W., Functional evaluation of the shoulder following latissimus dorsi muscle transfer, *Annals of Plastic Surgery* **35**(4) (1995), 349–355.

Goldner, R.D., Howson, M.P., Nunley, J.A., Fitch, R.D., Belding, N.R. and Urbaniak, J.R., One hundred eleven thumb amputations: Replantation vs revision, *Microsurgery* **11**(3) (1990), 243–250.

Harvey, P. and Gench, B., A comparison of static grip strength measurements taken on the Jamar dynamometer and the BTE [Abstract], *Journal of Hand Therapy* **6**(1) (1993), 53–54.

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.

King, J.W. and Berryhill, B.H., A comparison of two static grip testing methods and its clinical applications: A preliminary study, *Journal of Hand Therapy* **1** (1988), 204–208.

King, J.W. and Berryhill, B.H., Assessing maximum effort in upper extremity functional testing, *Work* **1**(3) (1988), 65–76.

Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.

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.

Rondinelli, R.D., Dunn, W., Hassanein, K.M., Keesling, C.A., Meredith, S.C., Schulz, T.L. and Lawrence, N.J., A simulation of hand impairments: Effects on upper extremity function and implications toward medical impairment rating and disability determination, *Archives of Physical Medicine & Rehabilitation* **78**(12) (1997), 1358–1363.

Wilke, N.A., Sheldahl, L.M., Dougherty, S.M., Levandoski, S.G. and Tristani, F.E., Baltimore Therapeutic Equipment Work Simulator: Energy expenditure of work activities in cardiac patients, *Archives of Physical Medicine & Rehabilitation* **74**(4) (1993), 419–424.

Wolf, L.D., Klein, L. and Cauldwell-Klein, E., Comparison of torque strength measurements on two evaluation devices, *Journal of Hand Therapy* **2** (1987), 24–27.

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.

Jay, M.A., Lamb, J.M., Watson, R.L. and Young, I.A., Sensitivity and specificity of the indicators of sincere effort of the EPIC Lift Capacity test on a previously injured population [Abstract], *Physical Therapy* **78**(5) (1998), S64.

Matheson, L., Mooney, V., Caiozzo, V., Jarvis, G., Pottinger, J., DeBerry, C., Backlund, K., Klein, K. and Antoni, J., Effect of instructions on isokinetic trunk strength testing variability, reliability, absolute value, and predictive validity, *Spine* **17**(8) (1992), 914–921.

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) (1993), 65–81.

Matheson, L.N., Mooney, V., Holmes, D., Leggett, S., Grant, J.E., Negri, S. and Holmes, B., A test to measure lift capacity of physically impaired adults. Part 2 - Reactivity in a patient sample, *Spine* **20**(19) (1995), 2130–2134.

Reyna, J.R., Leggett, S.H., Kenney, K., Holmes, B. and Mooney, V., The effect of lumbar belts on isolated lumbar muscle: Strength and dynamic capacity, *Spine* **20**(1) (1995), 68–73.

ERGOS Work Simulator

Cooke, C., Dusik, L.A., Menard, M.R., Fairburn, S.M. and Beach, G.N., Relationship of performance on the ERGOS work simulator to illness behaviour in a workers' compensation population with low back versus limb injury, *Journal of Occupational Medicine* **36**(7) (1994), 757–762.

Dusik, L.A., Menard, M.R., Cooke, C., Fairburn, S.M. and Beach, G.N., Concurrent validity of the ERGOS work simulator versus conventional functional capacity evaluation techniques in a workers' compensation population, *Journal of Occupational Medicine* **35**(8) (1993), 759–767.

King, P.M., Tuckwell, N. and Barrett, T.E., A critical review of functional capacity evaluations, *Physical Therapy* **78**(8) (1998), 852–866.

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) (1993), 65–81.

Simonsen, J.C., Coefficient of variation as a measure of subject effort, *Archives of Physical Medicine & Rehabilitation* **76**(6) (1995), 516–520.

Work Recovery, (undated), *ERGOS units 1–5*, Available from Work Recovery Pty. Ltd., Tucson, Arizona, USA.

Isernhagen Functional Capacity Evaluation

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

Isernhagen, S.J., Contemporary issues in functional capacity evaluation, in: *The comprehensive guide to work injury management*, S.J. Isernhagen, Ed., Gaithersburg, MD: Aspen, 1995, pp. 410–429.

Isernhagen Work Systems, *Reliability and validity of the Isernhagen Work systems Functional Capacity Evaluation*, Duluth, Ill: Isernhagen Work Systems, 1996.

King, P.M., Tuckwell, N. and Barrett, T.E., A critical review of functional capacity evaluations, *Physical Therapy* **78**(8) (1998), 852–866.

Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.

Key Method Functional Capacity Evaluation Assessment

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

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

King, P.M., Tuckwell, N. and Barrett, T.E., A critical review of functional capacity evaluations, *Physical Therapy* **78**(8) (1998), 852–866.

Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.

Lido WorkSET

Capodaglio, P., Gibellini, R., Grilli, C. and Bazzani, G., The assessment of functional capacity in workers with the thoracic outlet syndrome. A pilot study, [Article in Italian], *G Ital Med Lav Ergon* **19**(2) (1997), 15–19.

Ford, D., Kwak, A. and Wolfe, L.D., Grip strength decrease and recovery following isotonic exercise [Abstract], *Journal of Hand Therapy* **3**(1) (1990), 36.

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.

Wolf, L.D., Matheson, L.N., Ford, D.D. and Kwak, A.L., Relationships among grip strength, work capacity and recovery, *Journal of Occupational Rehabilitation* **6**(1) (1996), 57–70.

MESA/System 2000

Bordieri, J.E. and Musgrave, J., Client perceptions of the Microcomputer Evaluation and Screening Assessment, *Rehabilitation Counseling Bulletin* **32**(4) (1989), 342–345.

Janikowski, T.P., Berven, N.L. and Bordieri, J.E., Validity of the Microcomputer Evaluation Screening and Assessment aptitude scores, *Rehabilitation Counseling Bulletin* **35**(1) (1991), 38–51.

Janikowski, T.P., Bordieri, J.E. and Musgrave, J.R., Construct validation of the academic achievement and general educational development subtests of the Microcomputer Evaluation Screening and Assessment (MESA), *Vocational Evaluation & Work Adjustment Bulletin* **23**(1) (1990), 11–16.

Janikowski, T.P., Bordieri, J.E., Shelton, D. and Musgrave, J., Convergent and discriminant validity of the Microcomputer Evaluation Screening and Assessment (MESA) interest survey, *Rehabilitation Counseling Bulletin* **34**(2) (1990), 139–149.

Stoelting, C., A study of the construct validity of the MESA, *Vocational Evaluation & Work Adjustment Bulletin* **23**(3) (1990), 85–91.

Progressive Isoinertial Lifting Evaluation (**PILE**)

Curtis, L., Mayer, T.G. and Gatchel, R.J., Physical progress and residual impairment quantification after functional restoration. Part III: Isokinetic and isoinertial lifting capacity, *Spine* **19**(4) (1994), 401–405.

Hazard, R.G., Fenwick, J.W., Kalisch, S.M., Redmond, J., Reeves, V., Reid, S. and Frymoyer, J.W., Functional restoration with behavioural support: A one-year prospective study of patients with chronic low-back pain, *Spine* **14**(2) (1989), 157–161.

Hazard, R.G., Haugh, L.D., Green, P.A. and Jones, P.L., Chronic

low back pain: The relationship between patient satisfaction and pain, impairment and disability outcomes, *Spine* **19**(8) (1994), 881–887.

Mayer, T.G., Barnes, D., Nichols, G., Kishino, N.D., Coval, K., Piel, B., Hoshino, D. and Gatchel, R.J., Progressive isoinertial lifting evaluation II: A comparison with isokinetic lifting in a disabled chronic low-back pain industrial population, *Spine* **13**(9) (1988), 998–1002.

Mayer, T.G., Mooney, V., Gatchel, R.J., Barnes, D., Terry, A., Smith, S. and Mayer, H., Quantifying postoperative deficits of physical function following spinal surgery, *Clinical Orthopaedics & Related Research* **244** (1989), 147–157.

Rainville, J., Ahern, D.K., Phalen, L., Childs, L.A. and Sutherland, R., The association of pain with physical activities in chronic low back pain, *Spine* **17**(9) (1992), 1060–1064.

Polinsky Functional Capacity Assessment

Isernhagen, S., Role of functional capacities assessment after rehabilitation, in: *Ergonomics: The physiotherapist in the workplace*, M.I. Bullock, Ed., London: Churchill Livingstone, 1990, pp. 259–297.

Isernhagen, S.J., Mokros, K., Miller, M. and Johnson, L., Functional capacities assessment research: The relationship of age and gender to functional performance - Patients and uninjured subjects, in: *Work injury: Management and prevention*, S.J. Isernhagen, Ed., Gaithersburg, MD: Aspen, 1988, pp. 184–191.

Lechner, D., Work technology review (Polinsky Function Capacity Assessment), *Work* **2**(1) (1991), 70–71.

Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.

Piela, C.R., Hallenberg, K.K., Geoghegan, A.E., Monsein, M.R. and Lindgren, B.R., Prediction of functional capacities, *Work* **6**(2) (1996), 107–113.

Physical Work Performance Evaluation (**PWPE**)

Bevington, J., Warner, L., Hyde, S.D.A., Lechner, D.E. and Gossman, M.R., Performance values on four coordination tasks for healthy, working-aged adults [Abstract], *Physical Therapy* **74**(5) (1994), S99.

Buckley, E., Rasmussen, A.A., Lechner, D., Gossman, M.R., Quintana, J.B. and Grubbs, B., The effects of lumbosacral support belts and abdominal muscle strength on functional lifting ability in healthy women [Abstract], *Physical Therapy* **74**(5) (1994), S27.

King, P.M., Tuckwell, N. and Barrett, T.E., A critical review of functional capacity evaluations, *Physical Therapy* **78**(8) (1998), 852–866.

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.

Lechner, D.E., 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) (1993), S27.

Lechner, D.E., Sheffield, G.L., Page, J.J. and Jackson, J.R., Predictive validity of a functional capacity evaluation: The physical work performance evaluation [Abstract], *Physical Therapy* **76**(5) (1996), S81.

Prim, J.F., Shealy, S.A., Lechner, D.E., Gossman, M.R. and Bradley, E., Factors influencing the lifting ability of healthy females 20 to 35 years of age [Abstract], *Physical Therapy* **73**(6) (1993), S51.

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) (1993), 356–366.

Quantitative Functional Capacity Evaluation (**QFCE**)

Yeomans, S.G. and Liebenson, C., Functional capacity evaluation and chiropractic case management, *Topics in Clinical Chiropractic* **3**(3) (1996), 15–25.

Singer/New Concepts Vocational Evaluation System (**Singer VES**)

Gannaway, T.W. and Sink, J.M., The relationship between the vocational evaluation system by Singer and employment success in occupational groups, *Vocational Evaluation & Work Adjustment Bulletin* **11**(2) (1978), 38–45.

Gannaway, T.W., Sink, J.M. and Becket, W.C., A predictive validity study of a job sample program with handicapped and disadvantaged individuals, *Vocational Guidance Quarterly* **29**(1) (1980), 4–11.

Smith Physical Capacity Evaluation (**Smith PCE**)

Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.

Smith, S.L., Cunningham, S. and Weinberg, R., Predicting reemployment of the physically disabled worker, *Occupational Therapy Journal of Research* **3**(3) (1983), 178–179.

Smith, S.L., Cunningham, S. and Weinberg, R., The predictive validity of the functional capacities evaluation, *American Journal of Occupational Therapy* **40**(8) (1986), 564–567.

Spinal Function Sort

Browning, J., Juska, C., Howe, E., Mackie, H., Sevil, B. and Cusi, M.F., Relating critical physical job demands to ongoing gains in functional capacity for workers with back injuries, *Proceedings of the 2nd Annual Scientific Meeting of the Australasian Faculty of Rehabilitation Medicine* (pp. 159–165), Adelaide, SA: ACRM, 1994.

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) (1993), 15–30.

Sufka, A., Hauger, B., Trenary, M., Bishop, B., Hagen, A., Lozon, R. and Martens, B., Centralization of low back pain and perceived functional outcome, *Journal of Orthopaedic & Sports Physical Therapy* **27**(3) (1993), 205–212.

Valpar Component Work Samples (**Valpar CWS**)

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

Barry, P., Correlational study of a psychosocial rehabilitation program, *Vocational Evaluation & Work Adjustment Bulletin* **15** (1982), 112–117.

Bielecki, R.A. and Growick, B., Validation of the Valpar independent problem-solving work sample as a screening tool for brain damage, *Vocational Evaluation & Work Adjustment Bulletin* **17**(2) (1984), 59–61.

Cady, D.C., The correspondence of two vocational assessment devices in the prediction of job success, *Dissertation Abstracts International* **44**(1-B) (1983), 287.

Cederlund, R., The use of dexterity tests in hand rehabilitation, *Scandinavian Journal of Occupational Therapy* **2**(3–4) (1995), 99–104.

Dusik, L.A., Menard, M.R., Cooke, C., Fairburn, S.M. and Beach, G.N., Concurrent validity of the ERGOS work simulator versus conventional functional capacity evaluation techniques in a workers' compensation population, *Journal of Occupational Medicine* **35**(8) (1993), 759–767.

Growick, B., Kaliope, G. and Jones, C., Sample norms for the hearing-impaired on select components of the Valpar work sample series, *Vocational Evaluation & Work Adjustment Bulletin* **16**(2) (1983), 56–57, 68.

Jones, C. and Lasiter, C., Worker-non-worker differences on three Valpar component work samples, *Vocational Evaluation & Work Adjustment Bulletin* **10**(3) (1977), 23–27.

Kochevar, R.J., Kaplan, R.M. and Weisman, M., Financial and career losses due to rheumatoid arthritis: A pilot study, *Journal of Rheumatology* **24**(8) (1997), 1527–1530.

Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1997), 37–47.

Mott, J.H., Vocational screening tool for neurological impairment. VALPAR 6: Independent problem solving work sample [Abstract], *Dissertation Abstracts International* **54**(1-A) (1993), 157.

Rondinelli, R.D., Dunn, W., Hassanein, K.M., Keesling, C.A., Meredith, S.C., Schulz, T.L. and Lawrence, N.J., A simulation of hand impairments: Effects on upper extremity function and implications toward medical impairment rating and disability determination, *Archives of Physical Medicine & Rehabilitation* **78**(12) (1997), 1358–1363.

Saxon, J.P., Spitznagel, R.J. and Shellhorn-Schutt, P.K., Intercorrelations of selected VALPAR Component Work Samples and General Aptitude Test Battery scores, *Vocational Evaluation & Work Adjustment Bulletin* **16**(1) (1983), 20–23.

Schult, M., Söderback, I. and Jacobs, K., Swedish use and validation of Valpar work samples for patients with musculoskeletal neck and shoulder pain, *Work* **5**(3) (1995), 223–233.

Sen, S., Fraser, K., Evans, O.M. and Stuckey, R., A comparison of the physical demands of a specific job and those measured by standard functional capacity assessment tools, in: *Ergonomics and human environments: Proceedings of the 27th Annual Conference of the Ergonomics Society of Australia*, V. Propovic and M. Walker, Eds., Coolum, Qld: Ergonomics Society of Australia, 1991, pp. 263–268.

Valpar International Corporation, *Valpar Component Work Sample manual* (Work Samples 1–12, 15, 16 and 19). Tucson, Arizona: Valpar International Corporation, 1993.

WEST Standard Evaluation

Carlton, R.S., The effects of body mechanics instruction on work performance, *American Journal of Occupational Therapy* **41**(1) (1987), 16–20.

Dueker, J.A., Ritchie, S.M., Knox, T.J. and Rose, S.J., Isokinetic trunk testing and employment, *Journal of Occupational Medicine* **36**(1) (1994), 42–48.

Egeskov, R., *Select normative data of bilateral lifting capacity and the usage of the W.E.S.T. comprehensive weights system*, Unpublished Graduate Diploma in Occupational Health & Safety thesis, Queensland University of Technology, Brisbane, Qld, 1989.

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, The University of Sydney, Sydney, NSW, 1995.

Lechner, D., Roth, D. and Straaton, K., Functional capacity evaluation in work disability, *Work* **1**(3) (1991), 37–47.

Mayer, T.G., Gatchel, R.J., Kishino, N., Keeley, J., Capra, P., Mayer, H., Barnett, J. and Mooney, V., Objective assessment of spine function following industrial injury: A prospective study with comparison group and one-year follow-up, *Spine* **10**(6) (1985), 482–493.

Moran, M. and Strong, J., Outcomes of a rehabilitation programme for patients with chronic back pain, *British Journal of Occupational Therapy* **58**(10) (1995), 435–438.

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

Sen, S., Fraser, K., Evans, O.M. and Stuckey, R., A comparison of the physical demands of a specific job and those measured by standard functional capacity assessment tools, in: *Ergonomics and human environments: Proceedings of the 27th Annual Conference of the Ergonomics Society of Australia*, V. Propovic and M. Walker, Eds., Coolum, Qld: Ergonomics Society of Australia, 1991, pp. 263–268.

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, *Proceedings of the 19th National Conference of the Australian Association of Occupational Therapists – Volume 2* (pp. 245–251). Perth, WA: AAOT, 1997.

Velozo, C.A., Lustman, P.J., Cole, D.M., Montag, J.A. and Eubanks, B., Prediction of return to work by rehabilitation professionals, *Journal of Occupational Rehabilitation* **1**(4) (1991), 271–280.

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.

Wolf, L.D., Klein, L. and Cauldwell-Klein, E., Comparison of torque strength measurements on two evaluation devices, *Journal of Hand Therapy* **2** (1987), 24–27.

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

Ping, C.L.T.W., Keung, S.C.F. and Yee, P.L.W., Functional assessment of repetitive strain injuries: Two case studies, *Journal of Hand Therapy* **9**(4) (1996), 394–398.

WorkAbility Mark III

King, P.M., Tuckwell, N. and Barrett, T.E., A critical review of functional capacity evaluations, *Physical Therapy* **78**(8) (1998), 852–866.

Shervington, J. and Balla, J., Screening workplace capabilities for competitive employment: Report on workplace feedback, in: *Industrial engineering in occupational health: ANZMA seminars*, (Vol. 3, No. 1), J.M. Farrell, Ed., Melbourne, Vic.: Australia & New Zealand MODAPTS Association, 1994, pp. 31–65.

Shervington, J. and Balla, J., WorkAbility Mark III: Functional assessment of workplace capabilities, *Work* **7**(3) (1996), 191–202.

Work Box

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.

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

Assessment	Types of validity			
	Face/Content	Criterion-related	Construct	Screening
AME	**0**	**0**	**0** *5* (*pre/post treatment change - lifting*)	**0**
ARCON	**0**	**5** (ARCON & dual inclinometry - lumbar ROM) *5* (*ARCON & AMA impairment rating*)	**0** *5* (*pre/post treatment change - static lift, push, pull*)	**0**
AssessAbility	**1** (MTM)	**1** (MTM)	**0**	**0**
Blankenship FCE	**3** (DOT physical demands) *2* (*DOT physical demands*)	**0**	**0** *3* (*behavioural profile*) *5* (*compared LBP subjects with max & sub-max performance*)	**0**
BTE Work Simulator	**0** *2* (*DOT physical demands*)	**5** (#181, #701, #901 - compared VO_2 & HR in simulated & actual light, med. & heavy tasks) **5** (#122, #141, #171, #181, #191, #502, #701, #802, #901 - compared VO_2, HR & BP in simulated & actual tasks) **5** (#131, #171, #181 & arm cranking - VO_2 & HR) **5** (#162 & Jamar - different elbow positions) *3* (*#162 & Jamar - F & M*) *5* (*pron/sup - attachment no. not specified & WEST 4*) *5* (*#162 & Jamar - injured & uninjured hands*) *5* (*attachment no. not specified -impairment rating as predictor of functional loss*)	**0** *5* (*#162, #801 - compared exercise methods for UE injury*) *5* (*attachment nos. not specified - compared subjects with fibromyalgia, RA & no disorder*) *5* (*#131, #162, #302, #502 - compared replantation & revision of thumb amp.*) *5* (*#131, #171, #181 - compared F & M for VO_2 & HR*) *5* (*#802 & pron/sup - attachment no. not specified - compared 2 types of surgery for brachial plexus lesions*) *5* (*#171, #181, #191B, #802 - compared control & shoulder surgery groups*)	**0** *5* (*#302, #502, #503, #601, #701 - CV cut-offs*) *5* (*#162, #302, #502 - level of effort*)
Cal-FCP (includes EPIC & SFS)	**0**	**0** *5* (*EPIC & SFS - prediction of work capacity*)	**0** *5* (*EPIC & SFS - level of effort*)	**0**
DOT-RFC	**5** (DOT physical demands)	**0**	**5** (factor analysis establishing 4 major factors)	**0**
EPIC (PLC II was precursor of EPIC LC)	**0**	**5** (PLC II & Lido Lift) **5** (PLC & Lido Passive Back Machine) **5** (EPIC & ERGOS - human vs. computer instructions)	**5** (pre/post treatment change - LBP, 3 age groups) *5* (*effect of using lumbar belt on lifting*) *5* (*effect of age, resting HR, weight*)	**3** (indicators of sincere effort)
ERGOS Work Simulator	**2** (DOT physical demands) *2* (*DOT physical demands, NIOSH guidelines*)	**5** (EPIC & ERGOS - human vs. computer instructions) **5** (ERGOS & therapist evaluation, workshop tasks, VCWS)	**0** *5* (*compared subjects with LBP & LL injuries*) *5* (*compared CVs for different groups*)	**0**
Isernhagen FCE	**2, 2** (DOT physical demands) *2* (*DOT physical demands*)	**0**	**0** *3* (*RTW outcome*) *4* (*compared psychophysical & kinesiophysical lifts; injured & uninjured groups*)	**0**
Key FCA	**2** (DOT physical demands) *2* (*DOT physical demands*)	**0**	**0** *2* (*"Validity" profiles from database*) *3* (*RTW reinjury rate*)	**0**
Lido WorkSET	**0**	**0** *3, 5* (*#52 - isotonic & isometric strength as predictors of work capacity*)	**0** *2* (*pre/post treatment change - thoracic outlet syndrome*) *5* (*attachment no. not specified - compared CTD & healthy groups*)	**0**

Table 3 Continued

Assessment	Types of validity			
	Face/Content	Criterion-related	Construct	Screening
MESA/System 2000	**0** *5* (*client perceptions*)	**0**	**5** (convergent/divergent - MESA with DAT & TABE) **5** (convergent/divergent - MESA Interest Survey with USES Interest Survey) **5** (convergent/divergent - MESA with GATB & WAIS-R) **5** (convergent/divergent - MESA with GATB)	**0**
PILE	**0**	**0**	**0** *5* (*pre/post treatment change - LBP; correlation between PILE & Cybex Liftask*) *5* (*compared spinal surgery & normal groups*) *5* (*pre/post treatment change - LBP; compared working & non-working groups*) *5* (*pre/post treatment change - LBP*) *5* (*pre/post treatment change - LBP surgery & non-surgery*) *5* (*correlation between PILE & pain & disability*)	**0**
Polinsky FCA	**0** *2, 2* (*DOT physical demands*)	**0** *5* (*client ability to predict lifting & standing tolerance*)	**0** *3* (*compared F/M, 3 age groups, injured/uninjured*)	**0**
PWPE	**2** (DOT physical demands)	**3, 5** (PWPE & RTW level) **3** (PWPE & RTW level)	**0** *3* (*floor-waist lift & anthropometrics to predict safe lifting max.*) *3* (*coordination component; compared F/M, 4 age groups*) *3, 5* (*differences in lifting & use of LS belt*)	**0**
QFCE	**3** (DOT physical demands)	**0**	**0**	**0**
Singer/New Concepts VES	**0**	**5** (VES & jobs) **5** (VES & job placement)	**0**	**0**
Smith PCE	**0** *2* (*DOT physical demands*)	**5, 5** (PCE & RTW)	**0**	**0**
Spinal Function Sort (SFS)	**0**	**0**	**5** (convergent - SFS & PSEQ, SES, PDI, WRQ & VAS) *2* (*pre/post treatment change - injured workers*) *5* (*correlation between SFS & chronicity*) *5* (*correlation between SFS & Oswestry - LBP*)	**0**
Valpar CWS	**2** (all VCWS - DOT aptitudes, physical demands, temperaments) **4** (#19) *2* (*DOT physical demands*) *3* (*#8 - physical demands*)	**5** (VCWS #4, 5, 8, 9 & 11, therapist evaluation, workshop tasks & ERGOS)	**5** (convergent - #4, #6, #7, #8, #9, #10, #11 & GATB aptitude) *2* (*all VCWS - correlations with numerous other tests*) *3* (*#7, #9, #11 - compared workers & non-workers*) *3* (*#2, #3, #5, #6, #7, #9, #11 correlations with GATB*) *3* (*#6, #7, #8, #11 - compared hearing impaired & other groups*)	**5** (#6 - compared subjects with physical impairment, psychiatric disability & brain damage) *3* (*#6 - neurological impairment*)

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

Assessment	Types of validity			
	Face/Content	Criterion-related	Construct	Screening
AME	Unknown	Unknown	*Good* (pre/post treatment change - lifting)	Unknown
ARCON	Unknown	**Poor** (ARCON & dual inclinometry) *Poor* (ARCON & AMA impairment rating)	*Good* (pre/post treatment change - static lift, push, pull)	Unknown
AssessAbility	Unknown	Unknown	Unknown	Unknown
Blankenship FCE	Unknown	Unknown	*Unknown* (compared LBP subjects with max & sub-max performance)	Unknown
BTE Work Simulator	Unknown	**Moderate - good** (#181, #701, #901 - compared VO_2 & HR in simulated & actual light, med. & heavy tasks) **Moderate** (#122, #141, #171, #181, #191, #502, #701, #802, #901 - compared VO_2, HR & BP in simulated & actual tasks) **Fair** (#171); **Poor** (#131, #181) - compared with arm cranking - VO_2 & HR **Good** (#162 & Jamar) *Poor* (pron/sup - attachment no. not specified & WEST 4) *Good* (#162 & Jamar); *Good* (injured hands); *Moderate* (uninjured hands) *Poor* (impairment rating as predictor of functional loss)	*Moderate* (#162); *Poor* (#801) - compared exercise methods for UE injury) *Moderate* (differentiate between patient (fibromyalgia & RA) & healthy groups); *Poor* (differentiate between fibromyalgia & RA groups) *Unknown* (#131, #162, #302, #502 - compared replantation & revision of thumb amp.) *Poor* (*NS difference*) (#131, #171, #181) - compared F & M (weight adjusted) for VO_2 & HR *Poor* (*NS difference*) (#802 & pron/sup - attachment no. not specified - compared 2 types of surgery for brachial plexus lesions) *Unknown* (#171, #181, #191B, #802 - compared control & shoulder surgery groups)	*Unable to determine* (#302, #502, #503, #601, #701 - CV cutoffs) *Unknown* (#162, #302, #502 - level of effort)
Cal-FCP (includes EPIC & SFS)	Unknown	*Good* (EPIC & SFS - prediction of work capacity)	*Good* (EPIC & SFS - level of effort)	Unknown
DOT-RFC	**Moderate** (DOT physical demands)	Unknown	**Moderate** (factor analysis establishing 4 major factors)	Unknown
EPIC (PLC II was precursor of EPIC LC)	Unknown	**Unknown** (PLC II & Lido Lift) **Unknown** (PLC & Lido Passive Back Machine) **Unknown** (EPIC & ERGOS - human vs. computer instructions)	**Good** (pre/post treatment change) *Poor* (*NS difference*) (effect of using lumbar belt on lifting) *Good* (effect of age, resting HR, weight)	Unknown
ERGOS Work Simulator	Unknown	**Moderate - good** (EPIC & ERGOS - human vs. computer instructions) **Moderate** (ERGOS & overall Physical Activity determination); **Poor - good** (ERGOS & therapist evaluation, workshop tasks, VCWS)	*Moderate* (differentiation between subjects with LBP & LL injuries) *Poor* (*NS difference*) (differentiation between client groups on basis of CV)	Unknown
Isernhagen FCE	Unknown	Unknown	*Moderate* (compared psychophysical & kinesiophysical lifts) *Poor* (compared injured & uninjured groups)	Unknown
Key FCA	Unknown	Unknown	Unknown	Unknown
Lido WorkSET	Unknown	*Good* (#52) - isotonic strength as predictor of work capacity); *Poor* (isometric strength as predictor of work capacity	*Good* (attachment no. not specified - compared CTD & healthy groups)	Unknown
MESA/System 2000	*Moderate* (client perceptions)	Unknown	**Moderate** (convergent/divergent - MESA with DAT & TABE)	Unknown

Table 4 Continued

Assessment	Types of validity			
	Face/Content	Criterion-related	Construct	Screening
			Moderate (convergent/divergent - MESA Interest Survey with USES Interest Survey) **Moderate** (convergent/divergent - MESA with GATB & WAIS-R) **Poor** (convergent/divergent - MESA with GATB)	
PILE	Unknown	Unknown	*Good* (pre/post treatment change - LBP) *Good* (pre/post treatment change - LBP) *Good* (pre/post treatment change - LBP) *Good* (pre/post treatment change - LBP surgery & non-surgery) *Moderate* (compared spinal surgery & normal groups) *Poor* (compared working & non-working groups) *Poor* correlation between PILE & Cybex Liftask *Poor* correlation between PILE & pain & disability	Unknown
Polinsky FCA	Unknown	*Poor* (client ability to predict lifting & standing tolerance)	Unknown	Unknown
PWPE	Unknown	**Fair - moderate** (PWPE & RTW level)	*Moderate* (differences in lifting & use of LS belt)	Unknown
QFCE	Unknown	Unknown	Unknown	Unknown
Singer/New Concepts VES	Unknown	**Poor - Moderate** (VES & jobs) **Moderate** (VES & job placement)	Unknown	Unknown
Smith PCE	Unknown	**Moderate** (PCE & RTW)	Unknown	Unknown
Spinal Function Sort (SFS)	Unknown	Unknown	**Good** (convergent - SFS & PSEQ, SES, PDI, WRQ & VAS) *Moderate* (correlation between SFS & chronicity) *Moderate* (correlation between SFS & Oswestry - LBP)	Unknown
Valpar CWS	**Poor** (#19)	**Poor** (VCWS #4, 5, 8, 9 & 11 and ERGOS); **Moderate** (VCWS & therapist evaluation); **Moderate** (VCWS & workshop tasks) *Poor* (impairment rating as predictor of functional loss, using #1)	**Moderate** (#4); **Poor - moderate** (#8, #9); **Poor** (#6, #7, #10, #11) - convergent validity with GATB aptitude) *Moderate* (#6) - correlations with neuropsychological tests; *Poor* (*NS difference*) - comparison of workers & subjects with mental illness) *Moderate* (#4 - compared hand injured & healthy groups) *Good* (#8 - differentiate between sick/not sick listed); *Unknown* (#9) *Moderate* (#5 - change in earning capacity with RA)	**Moderate** (#6 - compared subjects with physical impairment, psychiatric disability & brain damage)
WEST Std Eval	**Poor**	**Poor - fair** (MHRWS & 3-D motion analysis) *Poor* (*NS difference*) (prediction for RTW) *Unknown* (WEST criterion measure & Lido trunk dynamometer & future work injury)	*Unknown* (norms for different occupational groups, injury types, F/M) *Unknown* (compared US & Aust. "norms" - considered to be concurrent V)	Unknown