Work 37 (2010) 309–320
DOI 10.3233/WOR-2010-1083
IOS Press

Simultaneous bilateral hand strength testing in a client population, Part I: Diagnostic, observational and subjective complaint correlates to consistency of effort

Darrell W. Schapmire[a,*], James D. St. James[b], Larry Feeler[c] and Joe Kleinkort[d]
[a]*X-RTS Software Products, Inc., Hopedale, IL, USA*
[b]*Millikin University, Decatur, IL, USA*
[c]*WorkSTEPS, Inc., Austin, TX, USA*
[d]*Joseph A. Kleinkort, PC, Trophy Club, TX, USA*

Received 16 April 2009
Accepted 11 June 2009

Abstract. *Objectives:* 1. To determine if scores on pain questionnaires and overt behaviors during a functional capacity evaluation (FCE) were related to variability between repeated measures during a hand strength assessment. 2. To determine if failure of statistically-based validity criteria, as proposed by Schapmire, St. James and Townsend et al. [26] is likely to be due to pain.
Participants: 200 consecutive clients presenting for an FCE.
Methods: Subjects filled out pain questionnaires, were observed for various behaviors and were administered the distraction-based hand strength assessment.
Results: Clients failing two or more of the statistically-based validity criteria had higher scores on most pain questionnaires, presented with a higher frequency of various pain behaviors ($p < 0.05$ and < 0.001, respectively), and had a lower rate of relevant surgeries ($p < 0.001$). There was no statistically significant difference in the number of failed validity criteria between this group of clients and for normal subjects feigning weakness in a controlled study ($p > 0.05$).
Conclusions: Pain does not reasonably explain the failure of the statistically-based validity criteria. The protocol is appropriate for use in a client population.

Keywords: Distraction-based testing, validity of effort, functional capacity evaluation (FCE) Sincerity of effort

1. Background

1.1. Purpose

The purpose of the study was twofold. First, it was our goal to determine if common clinical impressions and scores on four pain questionnaires were predictive of the classification of validity of effort, using a distraction-based protocol consisting in part of simultaneous testing of both hands, a method described by Schapmire, St. James et al. [26]. Niemeyer, Matheson and Carlton [24] believed that the assessment of validity of effort would be compromised in repeated measures protocols if the assessment involves the affected body part. They cited "pain" as the factor that would result in excessive differences between repeated measures, although no mention was made of any experimental results confirming this belief. So the second purpose of

*Address for correspondence: Darrell Schapmire, MS, X-RTS Software Products, Inc., P.O. Box 171, Hopedale, IL 61747, USA. Tel.: +1 309 449 5483; E-mail: ds@xrts.com.

this study was to determine if pain would indeed credibly explain a failure of the statistically-based validity criteria as described in Schapmire et al. [26]. Waddell, McCulloch, Kummel and Venner has previously defined "distraction-based testing" as "non-emotional, non-surprising and non-hurtful" [44].

Hirsch et al. [15] found that clients in the "high Waddell score group," (clients judged to have positive indicators for three or more categories of nonorganic back pain as described by Waddell [44]. Hirsch [15] found they tended to have lower physical output in terms of lumbar ranges of motion, torque and maximum velocities during B-200 dynamometry. Hirsch stated that the results of the biomechanical testing for this group of subjects could be affected by abnormal illness behavior and, therefore, the physical measurements for ranges of motion, torques and velocities might not accurately reflect organic pathology.

Menard et al. [20] had findings very similar to Hirsch's in a study of compensation subjects during a "comprehensive motor evaluation." Menard identified a "global" pattern of performance by back pain subjects from a "High Waddell" group. The pattern included smaller ranges of motion, torque and maximum velocities, as well as lower physical output for isometric elbow flexion, isometric knee extension, and static grip measurements on the Jamar Hand Dynamometer. Neither Hirsch nor Menard assessed the reproducibility of physical performance parameters which they studied.

Investigating clinical and psychological presentation in upper extremity clients, Himmelstein et al. [14] compared reports of pain in a working client population to work-disabled clients. It was reported that those who were not working had a higher incidence of "indeterminate" diagnoses, reported more pain, expressed more anger toward the employer, and had a greater psychological response to perceived pain.

In a forced choice study, exaggerated facial expressions were identified with an accuracy level "above chance," although the accuracy level reported was insufficient to be used as the sole basis for making definitive conclusions regarding a client's presentation [11]. Furthermore, the fact that not all the clinicians agreed with one another in all instances, the findings indicated that the ability to identify exaggerated facial expressions is not a science, but an intuitive exercise akin to "an art".

An article with the memorable title, "The Seriously Uninjured Hand," is widely cited in FCE reports listing supportive references for assessing effort during hand strength testing [39]. The "bell curve" concept (stronger grip strength in the mid-range of motion), proposed in this article to objectively classify effort, was based on the results from two subjects, one believed by Stokes to be cooperative, the other believed by him to be uncooperative. Another study found high agreement between various clinical impressions believed to predict "low effort" and the results of a computer-assisted test in which data relative to the bell curve and Rapid Exchange Grip (REG) testing were analyzed, Stokes [40]. Stokes also reported that a "low tech" version of the protocol, using a hydraulic hand dynamometer, was 84.2% sensitive to what was believed to be "low effort," but no analysis of clients who had no behaviors believed by Stokes to be predict REG and bell curve characteristics was conducted. Therefore, no information regarding the specificity of the alternate protocol and, hence, no assessment of its accuracy was provided.

One study [13] introduced the concept of REG testing, although no specifics were provided with regard to the standardization of the test with regard to rate of grip exchange. Some qualified success was found in REG testing by Joughin who reported sensitivity and specificity of 81% and 93%, respectively, in the classification of effort in a population of normal subjects [17]. However, the results were tempered by the finding of "poorer sensitivities and specificities [with REG]" when the method was used in a clinical setting. Two previous related studies investigated force-time characteristics of sincere effort and feigned weakness during grip testing. A study by Smith, Nelson, Sadoff and Sadoff [38] reported sensitivity of up to 100% and specificity of up to 95% for asymptomatic males and sensitivity of up to 93.5% and specificity of up to 97.8% for asymptomatic females. In a study which applied Smith et al.'s validity criteria to a population of 60 clients, sensitivity of up to 85.0% and specificity of up to 96.7% was reported for males and sensitivity of up to 83.3% and specificity of up to 100% was reported for females [2]. Shechtman, Sindhu and Davenport resumed the study of the force-time curve to classify effort, but there is yet to be a follow-up study validating the use of time-force data for that purpose [35].

One study reported that electromyography (EMG) analysis in conjunction with force analysis "has potential," in classifying sincerity of effort, but that actual sub-maximal force values are reproducible [23]. In a controlled study [16], hand strength was assessed in 11 normal subjects on six different sessions conducted over a 3- to 5-week period of time to determine if EMG amplitude and mean power frequency readings

during hand strength testing varied according to effort when comparing data from repeated measures. It was found that neither EMG variable differed significantly between sincere effort and feigned weakness sessions.

In a study of 80 clients with neurological conditions and 470 clients with head injuries [9]. This study did not assess physical performance, but, rather, conducted psychological, cognitive and perceptional testing. The title of the study, "Effort Has a Greater Effect on Test Scores than Severe Brain Injury in Compensation Patients" is, in itself, instructive. It was concluded, in part, "[E]ffort has such a large effect that, if not controlled, it literally inverts the group differences [on test scores] between severe versus very mild traumatic brain injury patients."

The authors have not identified any studies which unequivocally support the use of the most prevalent methods of classifying validity of effort during hand strength assessment, namely the coefficient of variation (CV), REG testing and various methods of assessing the "Bell-Shaped Curve." Many studies and literature reviews have found these methods to be inaccurate for classifying effort during a hand strength assessment [1,3,4,6–8,10,12,18,21,22,26–34,36,37,41–43,47].

2. Methods

2.1. Subjects

There were two groups of subjects in this study. One group consisted of 200 consecutive clients who had undergone an FCE which included a test for sincerity of effort during a hand strength protocol involving simultaneous use of both hands as described by Schapmire et al. [26]. All 200 were tested by the first author of this study. Data from seventy-five (75) additional subjects were also compiled from the results of an online version of the simultaneous bilateral hand strength test. The online tests were administered by 16 different therapists performing the assessment in various locations throughout the country. All of these subjects were receiving work-related injury or long term disability benefits. The Millikin University Institutional Review Board waived review of this study inasmuch as the data are derived from archived records of test results and no personal identifiers are used in the reporting of the results.

2.2. Classification of hand strength results

All subjects in this study underwent hand strength testing using a Jamar Dynamometer and a Baseline pinch gauge to measure the amount of force production. This distraction-based protocol includes a unilateral hand grip and pinch strength measurements and activities which require the person taking the test to generate force simultaneously in both hands as described by Schapmire et al. [26]. The protocol consists of a randomized order of 66 trials, analyzed for consistency of effort with the seven statistical validity criteria, listed in Table 1. The criteria classified sincerity of effort, also referred to as "consistency of effort," as follows:

1. All seven validity criteria are passed = valid effort.
2. One failed validity criterion = equivocal, or "gray zone" results.
3. Two or more failed validity criteria = invalid effort.

2.3. Pain scales

Prior to physical assessment, all subjects were asked to fill out a battery of written questionnaires. Each was asked to "rate your current level of disability with '0' representing no disability at all and '100' representing total disability". Additionally, each client was asked to "rate your chances of having a good recovery, with '0' representing no chance at all and '100' representing absolute certainty of having a good recovery". Finally, clients were asked to complete written instruments related to their symptomatic and functional status.

For the purpose of this study, the following questionnaires were selected for the cervical spine, shoulder and upper extremity clients:

1. 0–10+ Pain Rating Scale.
2. Visual Analog Scale (VAS).
3. Modified Somatic Perceptions Questionnaire [19].
4. Quantified Pain Drawing [25].

For the cervical spine and upper extremity clients, only the raw scores for the first three scales were considered in the statistical analyses. The Quantified Pain Drawing was originally developed to assess clients with lower back injury, as such, the scoring system recommended by Ohlund [25] was not used. Therefore, for the purposes of this study for only clients with cervical spine and/or upper extremity injuries, the Quantified Pain Drawing was classified by the evaluator as either "having a reasonable or anatomically plausible distri-

Table 1
Simultaneous bilateral validity criteria and related statistics [26]

Criterion	Frequency of violation during 100 sincere effort sessions specificity	Frequency of violation during 100 feigned weakness sessions / sensitivity
$\geqslant$ 5 CV's $\geqslant$ 15%	0 / 100%	70 / 70%
Mean of all CV's $\geqslant$ 9.75%	0 / 100%	77 / 77%
$\geqslant$ 5 changes $\geqslant$ 14% (comparing unilateral forces to bilateral forces)	0 / 100%	63 / 63%
Mean of all force changes $\geqslant$ 15% (comparing unilateral forces to bilateral forces)	1 / 99%	77 / 77%
Mean CV $\geqslant$ 10% for selected bilateral data sets	0 / 100%	73 / 73%
$\geqslant$ 2 CV's $\geqslant$ 20% for selected bilateral data sets	0 / 100%	67 / 67%
One CV $\geqslant$ 13% for Lateral Pinch during bilateral testing	4 / 96%	62 / 62%

These criteria were developed in a controlled study of normal subjects. In the study, all 100 subjects were tested two times. In one session, they gave a maximum effort. In another session, they were instructed to attempt to consistently feign weakness.

bution of symptoms" or "not having a reasonable or anatomically plausible distribution of symptoms" for clients with cervical spine and upper extremity injuries.

In addition to the clients with cervical spine and upper extremity injuries, other clients having diagnoses that were not anatomically related to the upper extremities were also tested and included as subjects in this study. Subjects in this group included clients with back pain and lumbar surgery, (hereafter referred to collectively as "low back clients"), and clients having diagnoses of "fibromyalgia," "chronic pain," or "chronic fatigue" (hereafter referred to collectively as "fibromyalgia clients"). In addition to the pain questionnaires previously mentioned, these particular clients were asked to fill out the following instruments:

1. Oswestry Low Back Inventory [5].
2. Inappropriate Symptoms Questionnaire (first five items only) [45].
3. Waddell Disability Index [46].

The Quantified Pain Drawing was scored according to the criterion suggested by Ohlund et al. [25] for all non-cervical spine and non-upper extremity clients reporting low back pain, including the fibromyalgia clients who universally reported experiencing back pain.

Although hand strength is not directly related to the diagnoses of some of the subjects in this study, a two-step process was used to identify those clients whose participation in a lifting assessment would conceivably be very limited, if not completely absent. Such a process is more time efficient than systematically – and unnecessarily – performing hand strength assessments on subjects whose diagnoses were unrelated to the upper extremities. This selection process was implemented to obtain sufficient information to classify validity of effort in the event the client prematurely terminated the *lifting* test either voluntarily or as the result of a behavioral presentation considered by the evaluator to be "unsafe for assessing lifting capacity." Therefore, the non-cervical spine and non-upper extremity clients whose scores surpassed three or more of the thresholds listed below were subsequently screened for possible non-physiologic hand strength in a "cursory screening," described immediately following this list of pain instruments:

1. 0–10+ Pain Rating Scale, score $\geqslant$ 7.
2. VAS, score $\geqslant$ 6.5 cm.
3. Modified Somatic Perceptions Questionnaire, score $\geqslant$ 13.
4. Quantified Pain Drawing, score $\geqslant$ 24.
5. Oswestry Low Back Inventory, score $\geqslant$ 50%.
6. Inappropriate Symptoms Questionnaire, score $\geqslant$ 3 (first five items only).
7. Waddell Disability Questionnaire, score $\geqslant$ 6.

The use of these cutoffs as part of a selection process was not based on a published study. Rather, they were based on more than a decade of anecdotal experience, believed to identify clients who are more likely to essentially refuse to participate in a lifting evaluation.

2.4. Cursory screening procedure, manual testing

The cursory screen for non-cervical spine and non-upper extremity clients who were selected for administration of the simultaneous bilateral protocol consisted of one isometric grip of 3–4 seconds on each hand in Position 2 on the Jamar Hand Dynamometer, with the clients being instructed to give a maximum effort. Each

client undergoing the cursory screen also performed one "explosive grip" in Position 2 for each hand in the manner described by Schapmire et al. [26]. The clients undergoing this screening process were administered the complete simultaneous bilateral hand strength protocol if they met any two of the following criteria:

1. Hand strength weakness on the "radicular side."
2. Sub-normal hand strengths in both hands.
3. Positive "explosive grip" in either hand that exceeded the corresponding static grip measurement by more than 10 pounds.

The following manual strength tests, with client-initiated force, were administered to all clients in this study:

1. Shoulder flexion.
2. Shoulder adduction and abduction.
3. Shoulder internal and external rotation.
4. Elbow flexion and extension.
5. Wrist flexion and extension.

2.5. Clinical impressions

In addition to the manual strength tests listed above, clients who had medical histories or subjective complaints involving the low back were administered manual strength tests for lower extremity strengths. If obvious regional weakness (also called "breakaway weakness," "give way weakness" or "cogwheeling") occurred on two or more of the manual strength tests, "cogwheeling" was noted on the data collection sheet. If facial affect, verbalization and reports of pain and dysfunction were considered to be "extreme" by the evaluator, this impression was noted on the data collection sheet as "overreaction."

3. Results

Due to the small sample size for the gray zone group (nine subjects), their data have been omitted from all statistical analyses in this manuscript.

3.1. Client demographics

In 15 instances, no precise date of injury could be identified secondary to conflicting medical records or significant differences between insurance company records and the client's statements. For these cases, the date of injury was treated as "missing data." Seven of these cases occurred during the testing of subjects who passed all validity criteria, seven during the testing of clients who failed two or more of the criteria and one for a client producing equivocal hand strength test results. Individual data related to "Time Since Injury" were rounded to the nearest 0.5 month. The mean time between injury and the hand strength testing was 18.2 months (SD = 16.1) for persons who failed none of the validity criteria. The mean time between injury and testing for those who failed two or more criteria was 17.7 months (SD = 15.8). Persons failing a single criterion were, on average, 8.7 months (SD = 5.7) post-injury.

Referring to Table 2, 83 of the 200 subjects (41.5%) passed all seven of the criteria. Two or more of the validity criteria were failed by one hundred eight (108), or 54.0% of all clients. Not listed in Table 2 the nine remaining nine (9) subjects, 4.5% of the entire client population, produced gray zone (equivocal) results, failing only one validity criterion.

Referring again to Table 2, Category 1, the frequency of upper extremity surgeries, inclusive of the shoulder, for the group passing all hand strength validity criteria was 35/83 (42.2%). This population of clients included some whose medical histories involved surgeries on the shoulder, elbow, forearm, wrist, hand or fingers. The frequency of surgeries for Category 1 clients who failed two or more validity criteria was 21/108 (19.4%). Thus, the frequency of surgical interventions for clients passing all criteria was 2.2 times the frequency of subjects who failed two or more validity criteria. This group difference is statistically significant, χ^2 (1) = 11.70, $p = 0.001$.

Still referring to Table 2, there are no statistically significant differences in the frequencies for clients in Categories 2–8. The range for the χ^2 values for these categories range from 0.24–1.95, with p values ranging from 0.164–0.874. A total of 16 clients fell into Categories 9 and 10. These low back and lumbar surgery clients universally failed two or more hand strength validity criteria. Group differences in the frequency of clients in these diagnostic categories are statistically significant, χ^2 (1) = 5.58, $p = 0.018$.

3.2. Accuracy of clinical impressions

Table 3 reports the agreement between three clinical impressions and the bilateral hand test outcome. These impressions were related to over-reaction, cogwheeling and a judgment as to whether the distribution of symptoms on the Quantified Pain Drawing were reasonable. In all three cases, χ^2 test results show statistically

Table 2
Test outcome per diagnostic category

Diagnosis	Passed All validity criteria, N = 83	Failed ⩾ 2 validity criteria, N = 108	χ^2 and p values
Category 1: One or more upper extremity surgeries involving the elbow, forearm, hands or fingers (includes shoulder clients who also had surgeries on these parts of the body)	35 (42.2%)	21 (19.4%)	χ^2 (1) = 11.70 p 0.001
Category 2: One or more cervical spine surgeries plus one or more upper extremities surgeries involving the elbow, forearm hands or fingers	1 (1.2%)	3 (2.8%)	χ^2 (1) = 0.57 p = 0.452
Category 3: One or more cervical spine surgeries or confirmed cervical HNP (with radiculopathy)	7 (8.4%)	4 (3.7%)	χ^2 (1) = 1.95 p = 0.164
Category 4: Any shoulder surgery as primary diagnosis (does not include clients with cervical spine or other upper extremity surgeries)	15 (18.2%)	23 (21.3%)	χ^2 (1) = 0.31 p = 0.580
Category 5: Fracture in arm, wrist or hand (no history of upper extremity surgery)	2 (2.4%)	3 (2.8%)	χ^2 (1) = 0.03 p = 0.874
Category 6: Non-surgical clients reporting pain in at least one of the following areas: one or both upper extremities, one or both shoulders, cervical spine pain, cervical spine degenerative disc disease, cervical disc bulge	16 (19.3%)	23 (21.3%)	χ^2 (1) = 0.12 p = 0.731
Category 7: Diagnosis of at least one of the following: Fibromyalgia, chronic fatigue, chronic pain	3 (3.6%)	8 (7.4%)	χ^2 (1) = 1.24 p = 0.265
Category 8: Miscellaneous	4[1] (4.8%)	7[2] (6.5%)	χ^2 (1) = 0.24 p = 0.625
Category 9: Low back pain	0	9 (8.3%)	χ^2(1) = 7.26 p = 0.007
Category 10: Lumbar surgery	0	7 (6.5%)	χ^2 (1) = 5.58 P = 0.018

[1] Primary diagnoses: T4 fracture, cranial laceration, brachial stretch injury, widespread 3^{rd} degree burns (multiple skin grafts to shoulder and upper quadrant).
[2] Primary diagnoses: Rib resection (9^{th} and 10^{th}), cranial contusion (disputed loss of consciousness), T7 fracture, thoracic outlet syndrome, tarsal tunnel release, osteoarthristis with spurring on the thumb, lower extremity pain.

Table 3
Agreement between three clinical impressions and test classification

	Passed all validity criteria	Failed ⩾ 2 validity criteria	χ^2 and p
Was the client over-reactive (facial expression, verbalization)?	"Yes" for 5/83 (6.0%)	"Yes" for 65/108 (60.2%)	χ^2 (1) = 59.26 p = 0.000
Did the client cogwheel during manual strength testing?	"Yes" for 2/81 (2.4%)[1]	"Yes" for 39/108 (36.1%)	χ^2 (1) = 30.84 p = 0.000
Was the distribution of symptoms on the Pain Drawing anatomically plausible?[2]	N = 61 "Yes" for 54/61 (88.5%)	N = 70 "Yes" for 52/70 (74.2%)	χ^2 (1) = 4.28 p = 0.039

[1] Not assessed for two subjects who were referred for hand strength assessment only.
[2] "Reasonableness" for Quantified Pain Drawing was not assigned a numerical score since the instrument's original scoring instructions applied only to low back pain clients. Therefore when subjects had primary complaints related to the upper quadrant, upper extremities, head, face or lower extremities, a subjective assessment of the "reasonableness" of the distribution of symptoms was attempted.

significant differences between clients who passed all hand strength assessment validity criteria as compared to those who failed two or more criteria. Although the clients who failed two or more criteria had a higher frequency for all three impressions, nearly 40% of those who failed two or more criteria were not judged to be over-reactive, only 36.1% were believed to cogwheel during manual strength testing, and 74.2% appeared to report their symptoms in an anatomically plausible distribution on the Quantified Pain Drawing.

Table 4 reports the scores on four pain instruments: 0–10+ Pain Scale, VAS, Modified Somatic Perceptions, and Quantified Pain Drawing, scored as described by Ohlund [25]. Not all subjects chose to complete all the written pain questionnaires. Written instruments not completed by the clients were omitted from the statistical analyses, with the exception of the Oswestry Low Back Inventory which has a scoring system that does not require responses to all 10 items to be scored as described by Fairbanks [5].

Although there are statistically significant differences between group scores for all scales in Table 4 except the numeric score for the Quantified Pain Drawing, a focus on "statistical significance" is not advised.

Table 4
Agreement between client-reported pain and disability scores and classification of effort

Score on pain and disability scales	Passed all validity criteria (N, Mean, SD and Range)	Failed $\geqslant$ 2 validity criteria (N, Mean, SD and Range)	t -test results
0–10+ Pain Rating	N = 75 Mean = 4.51 SD = 2.36 Range = 0–10	N = 99 Mean = 5.92 SD = 2.63 Range = 0–10	$t = 3.66$ (172), $p = 0.000$
Visual Analog Score in Centimeters	N = 73 Mean = 4.95 SD = 3.85 Range = 0–10	N = 95 Mean = 6.06 SD = 2.76 Range = 0–10	$t = 2.55$ (166), $p = 0.012$
Modified Somatic Perceptions Score	N = 68 Mean = 7.24 SD = 6.55 Range = 0–28	N = 95 Mean = 10.66 SD = 6.97 Range = 0–30	$t = 3.16$ (161), $p = 0.002$
Quantified Pain Drawing Score[1]	N = 10 Mean = 38.90 SD = 23.90 Range = 1–65	N = 30 Mean = 26.93 SD = 19.73 Range = 3–90	$t = 1.58$ (38), $p = 0.123$
Client's self-reported rate of disability	N = 56 Mean = 56.93 SD = 26.53 Range = 0–100	N = 74 Mean = 69.91 SD = 22.83 Range = 2–100	$t = 2.99$ (128), $p = 0.003$
Client's self-reported chances of having a "good recovery"	N = 52 Mean = 54.73 SD = 34.82 Range = 0–100	N = 70 Mean = 40.37 SD = 32.21 Range = 0–100	$t = 2.28$ (120), $p = .025$

[1]Scored per Ohlund [25] if client reported low back pain as a source of pain and/or dysfunction. For most patients in this group, back pain was incidental to the primary complaint or diagnosis. The score in such cases refers to the number of squares on the grid that were marked by the client.

There is a complete overlap between the lower and upper ranges for all variables in Table 4, with the exception of the Modified Somatic Perceptions Questionnaire and the Quantified Pain Drawing. Group membership, thus, is not predicted by individual scores on these instruments.

Attention is called to the 27 clients in Table 4 in Categories 8–10. With the exception of four clients in these groups (one client with bone spurring in a thumb, one with thoracic outlet syndrome, one with a brachial stretch injury, and one with significant burns on the shoulder, arm and upper quadrant), the remaining clients have diagnoses that are not directly related to the upper extremity. However, they were identified during the cursory screen as individuals who would be likely to have limited participation in a lifting assessment. Of the clients so identified and tested, 23/27 (85.1%) failed two or more of the hand strength validity criteria.

Table 5 compares the number of failed criteria for eight different groups of subjects. This table also provides the "predicted range of scores for 95%" of each of six categories of clients, assuming a normal distribution of scores. This range is comprised of all scores falling $\pm$ 2 SD from the mean score for each group. Only marginal differences are seen when comparing the uppermost and lowermost predicted scores for all eight categories of subjects in Table 5.

In Table 5, the t-test values comparing the mean number of failed criteria for Category 1 subjects (normal subjects instructed to attempt to consistently feign weakness) to the means for clients for Categories 2-8 are found in Table 5. All t values fell below 1.0 with the exception of clients in Category 5, comprised of low back pain and low back surgery clients, clients diagnosed with fibromyalgia, and five clients whose diagnosis is not considered by the authors to be plausibly related to the upper extremities. Otherwise, t values ranged from 0.04 to 0.87. Diagnoses for these subjects are listed beneath the table. None of the t values are statistically significant, with all p values > 0.05.

Most noteworthy of the comparisons in Table 5 is the comparison of the distributions for failed criteria in Category 1 subjects, normal subjects who were instructed to feign weakness in Schapmire [26], to Category 8 subjects, consecutive clients tested independently by 16 different therapists using an online version of the test. The average number of failed criteria for Category 1 subjects was 4.89, SD = 1.85, as compared

Table 5

Distributions for failed validity criteria per diagnostic categories: Comparing results of a controlled study of normal subjects to clients who failed two or more criteria

Category 1: Normal subjects during instructed-noncompliant sessions	Category 2: Non-surgical cervical spine, shoulder or upper extremity pain	Category 3: Upper extremity fracture, or surgery on cervical spine, shoulder, upper extremity, or cervical HNP with confirmed radiculopathy, plus two clients with plausible[1] miscellaneous diagnoses	Category 4: Any shoulder surgery as primary diagnosis, not included in any other category of clients	Category 5: Low back pain or surgery, fibromyalgia and five clients with diagnoses not plausibly related to the upper extremity[2]	Category 6: All clients from Categories 2, 3 and 4	Category 7: All clients in this study, exclusive of those tested with the online test	Category 8: Consecutive clients, tested independently by 16 different therapists using an online version of the test
N = 100	N = 23	N = 33	N = 23	N = 29	N = 79	N = 108	N = 75
Mean = 4.89	Mean = 5.17	Mean 4.75	Mean = 5.03	Mean = 5.48	Mean = 5.04	Mean = 5.11	Mean = 4.90
SD = 1.92	SD = 1.85	SD = 1.99	SD = 1.70	PR = 1.63–8.53[3]	SD = 1.75	SD = 1.72	SD = 1.69
PR = 1.05–8.73[3]	PR = 1.47–8.87[3]	PR = 0.77–8.73[3]	SD = 1.63–8.53[3]	PR = 2.22–8.74[3]	PR = 1.57–8.54[3]	PR = 1.62–8.55[3]	PR = 1.52–8.28[3]
	t= 0.63[4]	t = 0.36[4]	t = 0.32[4]	t = 1.50[4]	t = 0.54[4]	t = 0.87[4]	t = 0.04[4]
	$p > 0.05$[4]	$p > 0.05$[4]	$p > 0.05$[4]	$p > 0.05$[4]	$p > 0.05$[4]	r $p > 0.05$[4]	$p > 0.05$[4]

[1]One thoracic outlet patient and one patient with osteoarthristis and bone spurring in the thumb.

[2]Includes all Category 7, 8 and 9 patients from Table 2, excluding one patient with osteoarthritis and bone spurring in the thumb and one patient with thoracic outlet syndrome (whose data were included in Category 3 because these diagnoses are plausibly related to the upper extremity). Included in Category 5 patients in this table are one patient with rib resection (9^{th} and 10^{th}), cranial contusion (disputed loss of consciousness) T7 fracture, tarsal tunnel release and lower extremity pain (diagnoses not plausibly related to the hands).

[3]PR is the predicted range for 95% of the population, comprising all scores $\pm$ 2 SD from the mean score for each patient category, assuming a normal distribution of scores.[4]Denotes t-Test results which compare means of the various patient groups to the mean number of failed criteria during feigned weakness session in a controlled study of normal subjects (Category 1). In all instances, the comparisons show non-significant differences.

to Category 8 clients (mean = 4.90, SD = 1.69). The t value in comparing the means was 0.04, $p > 0.05$. These results indicate there is no difference between the average number of failed criteria for these groups.

4. Discussion

There are group differences on the scores of most of the written pain instruments between the clients who passed all the hand strength validity criteria and those who failed two or more. Those who failed the hand strength validity criteria, as a group, tended to have pain drawings that were classified as "not reasonable or anatomically plausible." Similar results were obtained from questions related to client-perceived rates of disability and "chances of having a good recovery." It is emphasized, though, that these are only group differences. Group differences do not predict individual outcomes.

Subjects in this study who passed the validity criteria for the hand strength test were rarely judged to have exhibited extreme over-reaction in terms of facial expression, grimacing or groaning, and rarely presented with regional weakness during manual strength testing. Individuals who failed the validity criteria had a much higher incidence of such behaviors. However, many subjects who failed two or more validity criteria were not judged to be over-reactive. Since subjective impressions can not be standardized between observers, they should not be the primary basis for deferring an assessment of effort or as the sole basis for making predictions related to compliance during a test.

Given the lack of agreement between various impressions such as the ones investigated in this study, it may be tempting to argue that the solution is to "become better" at interpreting various phenomena such as facial affect. This process would presumably involve attempting to "fine tune" one's ability to more or less divine the presence or absence of exaggerated expressions of pain. Such an attempt also overlooks the very real possibility that many clients whose behavior is judged to be "unremarkable" may be just exactly that – unremarkable. Unremarkable presentations do not necessarily predict cooperation during a test. Furthermore, it is not readily apparent to the authors how it would be possible to standardize "interpretation" of observational data.

It is not possible to predict test outcome for individuals, based on the various scores for the scales investigated in this study, or on the presence or absence of the impressions investigated in this study. However, when clients presenting for FCE's have a cluster of features including high scores on pain questionnaires, cog-wheeling during manual strength testing, extreme overt pain behaviors, or produce questionable results during a cursory hand strength assessment as described herein, a complete assessment of sincerity of effort is appropriate to at least rule out the presence of non-cooperation. Conversely, the absence of such behaviors does not predict a "clean bill of health" with regard to cooperation during the hand strength assessment.

Those subjects who failed two or more of the validity criteria during hand strength assessment had a much *lower* rate of surgical interventions involving the cervical spine, shoulders and upper extremities than clients who passed the validity criteria (Table 2). But as a group they had higher scores on written pain instruments (Table 4). Furthermore, there were no differences between the two groups with regard to the frequency of clients in Categories 3–8 in Table 2. To conclude that persons who failed the validity criteria during simultaneous bilateral testing becaue of "pain," we must also believe that those who passed the criteria experienced less pain, even though they had a much higher rate of surgical intervention as a group. We would also have to believe that those who failed the assessment of validity had somehow been victimized by substandard care and under diagnosis, thus accounting for the lower rate of surgical intervention for that group. Furthermore, we would have to contend that those who passed the validity testing may have had unwarranted surgeries, but that their surgical procedures did not result in significant pain during the test. Most notably, we would have to completely ignore the fact that 16 of the subjects who failed simultaneous bilateral testing of the hands had diagnoses of low back pain, low back surgery or lower extremity complaints, none of which are related to upper extremity function – *and yet, as a group, they failed more validity criteria than any other client group*. Thus, for the clients in this study, "pain" is not a reasonable excuse for the failure to perform consistently during a test that involves the hands.

Normally, it is an inefficient use of a clinician's time to administer a hand strength assessment to clients who have diagnoses unrelated to the upper extremities. However, twenty-three (23) subjects in this study (Categories 8–10 in Table 2) had diagnoses related to the low back, or had miscellaneous diagnoses that have no direct bearing on upper extremity function – and overwhelmingly, the subjects in these categories failed two or more of the hand strength validity criteria. These

clients were selected for test administration on the basis of the previously-described cursory screen. The information presented herein with regard to the inaccuracy of clinical impressions and the inability of pain questionnaires to predict test outcome highlights the importance of knowing when it is completely appropriate to assess uninvolved parts of the body to assess validity of effort.

For every group of clients in this study, the mean number of failed criteria, the SD's of the various distributions of scores for these subjects, and the frequency of equivocal test results for the entire client population were nearly identical to the findings reported by Schapmire [26]. In fact, there are no statistically significant differences between Category 1 subjects in Table 5 (normal subjects instructed to feign weakness) and any of the client groups in Categories 2–8. Furthermore, the upper and lower ranges of the statistically-predicted distributions of scores for Categories 2–8 in Table 5 fall between the upper and lower limits of the predicted range for the normal subjects who feigned weakness in a controlled study. Lastly, the frequency of gray zone tests in this study (4.5%) is nearly identical to the frequency of such tests in the controlled study described by Schapmire [26]. These analyses demonstrate that the validity criteria, in fact, did not penalize clients by causing them to fail validity criteria at a higher frequency than was observed under experimental circumstances when normal subjects were instructed to attempt to feign weakness.

Conventional wisdom for many years has been that pain affects test performance to the extent that assessments of validity of effort are not appropriate if the testing involves the injured body part. According to Niemeyer [24], validity of effort testing must be limited to the testing of uninvolved parts of the body in a client population and can not be used to assess an involved body part. The authors of this study reject that belief in light of the findings presented herein.

In addition to rejecting the concept that "pain" will result in an increased number of failed criteria for clients, the authors also reject the notion that the length of time off work may somehow affect test performance. The clients who failed two or more validity criteria in this study were actually off work for a slightly shorter period of time than those subjects who passed all the validity criteria.

Regarding possible weaknesses of this study, the client data, with the exception of the online test data, was collected by the first author, raising the possible issue that other evaluators would have obtained different results. However, in the controlled study described in Schapmire [26], multiple evaluators collected data on an independent basis and all had the same result, whether testing normal subjects who were giving a good effort or normal subjects who were feigning weakness. With only one error in test classification for 200 sets of data during a controlled study – combined with the first-author-to-online-test comparison in this study – the protocol appears to have similar results across testers. Finally, it is also noted that Category 8 clients (Table 5) who failed two or more validity criteria – and were tested independently by multiple therapists with the online version of the test. These clients had an average number of 4.90 failed validity criteria. This average was nearly identical to the mean number of failed criteria (4.89) for Category 1 subjects who were feigning weakness in a controlled study conducted by the first two authors of this study.

One of the strengths of this study is that the information constitutes new information which has practical application. Specifically, this study pertains to simultaneous bilateral testing of the hands, a "distraction-based" testing method as defined by Waddell [44], i.e., tests that are "non-emotional, non-surprising, and non-hurtful." The "distraction" in the hand strength assessment is the simultaneous testing of both hands. This study demonstrates that impressions regarding "over-reaction," whether or not a pain drawing is "reasonable," and whether or not the client "cogwheeled" during manual strength testing are all subjective judgments that do not necessarily predict test outcome when a uniform analysis of variability is applied to physical performance data. At the same time, it advances the idea that test behavior, specifically the degree of consistency of effort, can be measured with a statistical analysis. Unlike categorical data such as "impressions," a standardized statistical analysis allows for the formulation of a reasonable hypothesis regarding the cause-and-effect relationship between behavior and test outcome.

5. Conclusions

The research hypotheses are rejected. Clinical impressions and scores on pain questionnaires do not predict classification of validity of effort during the simultaneous bilateral hand strength assessment. Given the significantly lower number of surgical interventions for clients failing the hand strength validity criteria, "pain" does not appear to be a reasonable explanation as to why clients fail the validity criteria proposed by Schap-

mire [26]. The number of failed criteria reported by Schapmire [26] for subjects who failed two or more criteria is nearly identical to the number of failed criteria for all such test results when the test is administered by other individuals. Given these facts, the simultaneous bilateral hand strength protocol appears to be well-suited for use in identifying abnormal test behaviors in the clinical setting.

References

[1] R.F. Ashford, S. Nagelburg and, R. Adkins, Sensitivity of the Jamar Dynamometer in detecting submaximal grip effort, *J Hand Surg [Am]* **3** (1996), 402–405.

[2] S.N. Chengular, G.A. Smith, R.C. Nelson and A.M. Sadoff, Assessing sincerity of effort in maximal grip strength tests, *Am J Phys Med Rehabil* **69** (1990), 148–153.

[3] L. De Smet and J. Londers, Repeated grip strength at one month interval and detection of voluntary submaximal effort, *Acta Orthopaedica Belgica* **69** (2003), 142–144.

[4] Z. Dvir, Coefficient of variation in maximal and feigned static and dynamic grip efforts, *Am J Phys Med Rehabil* **78** (1999), 216–221.

[5] C. Fairbanks, J. Couper, J. Davies and J. O'Brien, The Oswestry low-back pain disability questionnaire, *Physiotherapy* **66** (1980), 271–273.

[6] A.H. Fairfax, R. Balnave and R.D. Adams, Variability of grip strength during isometric contraction, *Ergonomics* **38** (1995), 1819–1830.

[7] D.A. Fishbain, R. Cutler, H.L. Rosomoff and R.S. Rosomoff, Chronic pain disability exaggeration/malingering and submaximal effort research, *Clin J Pain* **15** (1999), 244–274.

[8] S. Goldman, T.D. Cahalan and K.N. An, The injured upper extremity and the JAMAR five-handle position grip test, *Am J Phys Med Rehabil* **70** (1991), 306–308.

[9] P. Green, M.L. Rohling, P.R. Lees-Haley and L.M. Allen, Effort has a greater effect on test scores than severe brain injury in compensation claimants, *Brain Injury* **15** (2001), 1045–1060.

[10] Z. Gutierrez and O. Shechtman, The effectiveness of the five-handle position grip strength test in detecting sincerity of effort in men and women, *Am J Phys Med Rehabil* **82** (2003), 847–855.

[11] H. Hadjistavropoulos, K. Craig, T. Hadjistavropoulos and G. Poole, Subjective judgments of deception in pain expression: accuracy and errors, *Pain* **65** (1996), 251–258.

[12] A. Hamilton, R. Balnave and R. Adams, Grip strength testing reliability, *J Han Ther* **7** (1994), 163–170.

[13] D.H. Hildreth, W.C. Breidenbach, G.D. Lister and A.D. Hodges, Detection of submaximal effort by use of the rapid exchange grip, *J Hand Surg [Am]* **14** (1989), 742–745.

[14] J. Himmelstein, M. Feuerstein, E. Stanek et al., Work-related upper-extremity disorders and work disability: clinical and psychosocial presentation, *J Occup Environ Med* **11** (1995), 1278–1286.

[15] G. Hirsch, G. Beach, C. Cooke et al., Relationship between performance on lumbar dynamometry and Waddell score in a population with low-back pain, *Spine* **16** (1991), 1039–1043.

[16] E. Hoffmaster, R. Lech and B.R. Niebuhr, Consistency of sincere and feigned grip exertions with repeated testing, *J Occup Med* **35** (1993), 788–794.

[17] K. Joughin, P. Gulati, S.E. Mackinnon et al., An evaluation of rapid exchange and simultaneous grip tests, *J Hand Surg [Am]* **18** (1993), 245–252.

[18] D.E. Lechner, S.F. Bradbury and L.A. Bradley, Detecting sincerity of effort: a summary of methods and approaches, *Phys Ther* **78** (1998), 867–888.

[19] C. Main, The Modified Somatic Perception Questionnaire (MSPQ), *J Psychosom Res* **6** (1983), 503–513.

[20] M.R. Menard, C. Cooke, S.R. Locke et al., Pattern of performance in workers with low back pain during a comprehensive motor performance evaluation, *Spine* **12** (1994), 1359–1366.

[21] B.R. Niebuhr and R. Marion, Detecting sincerity of effort when measuring grip strength, *Am J Phys Med* **66** (1987), 16–24.

[22] B.R. Niebuhr and R. Marion, Voluntary control of submaximal grip strength, *Am J Phys Med Rehabil* **69** (1990), 96–101.

[23] B.R. Niebuhr. R. Marion and S.M Hasson, Electromyographic analysis of effort in grip strength assessment, *Electromyogr Clin Neurophysiol* **33** (1993), 149–156.

[24] L.O. Niemeyer, L.N. Matheson and R.S. Carlton, Testing consistency of effort: BTE Work Simulator, *Industrial Rehabilitation Quarterly* **2** (1989).

[25] C. Ohlund, C. Eek, S. Palmblad et al., Quantified pain drawing in subacute low back pain, *Spine* **21** (1996), 1021–1031.

[26] D. Schapmire, J.D. St. James, R. Townsend et al., Simultaneous bilateral testing: validation of a new protocol to detect insincere effort during grip and pinch strength testing, *J Hand Ther* **15** (2002), 242–250.

[27] O. Shechtman, Is the coefficient of variation a valid measure for detecting sincerity of effort of grip strength? *Work* **13** (1999), 163–169.

[28] O. Shechtman, Using the coefficient of variation to detect sincerity of effort of grip strength: A literature review, *J Hand Ther* **13** (2000), 25–32.

[29] O. Shechtman and C. Taylor, The use of the rapid exchange grip test in detecting sincerity of effort. Part II: The validity of the rapid exchange grip test, *J Hand Ther* **13** (2000), 203–210.

[30] O. Shechtman, The coefficient of variation as a measure of sincerity of effort of grip strength. Part I: The statistical principle, *J Hand Ther* **14** (2001), 180–187.

[31] O. Shechtman, The coefficient of variation as a measure of sincerity of effort of grip strength, Part II: sensitivity and specificity, *J Hand Ther* **14** (2001), 188–194.

[32] O. Shechtman and C. Taylor, How do therapists administer the rapid exchange grip test? a survey, *J Hand Ther* **15** (2002), 53–61.

[33] O. Shechtman, Z. Gutierrez and E. Kokendofer, Analysis of methods used to detect submaximal effort with the five-rung grip strength test, *J Hand Ther* **18** (2005), 10–18.

[34] O. Shechtman., S. Anton, W.F. Kanasky and M.E. Robinson, The use of the coefficient of variation in detecting sincerity of effort: a meta-analysis, *Work* **16** (2006), 335–341.

[35] O. Shechtman, B.S. Sindhu and P.W. Davenport, Using the force-time curve to detect maximal grip strength effort, *J Hand Ther* **20** (2007), 37–47.

[36] O. Shechtman and S.K. Goodall, The administration and interpretation of the rapid exchange grip test: a national survey, *J Hand Ther* **21** (2008), 18–27.

[37] J.C. Simonsen, Coefficient of variation as a measure of subject effort, *Arch Phys Med Rehabil* **76** (1995), 516–520.

[38] G.A. Smith, R.C. Nelson, S.J. Sadoff and A.M. Sadoff, Assessing sincerity of effort in maximal grip strength tests, *Am J Phys Med Rehabil* **68** (1989), 73–80.

[39] H.M. Stokes, The seriously uninjured hand – weakness of grip, *J Occup Med* **25** (1983), 683–684.
[40] H.M. Stokes, K.W. Landrieu, B. Domangue and S. Kunen, Identification of low-effort patients through dynamometry, *J Hand Surg [Am]* **20** (1995), 1047–1056.
[41] C. Taylor and O. Shechtman, The use of the rapid exchange grip test in detecting sincerity of effort. Part I: The administration of the rapid exchange grip test, *J Hand Ther* **13** (2000), 195–202.
[42] M.W. Tredgett, L.J. Pimble and T.R. Davis, The detection of feigned hand weakness using the five position grip strength test, *J Hand Surg [Br]* **24** (1999), 426–428.
[43] M.W. Tredgett and T.R. Davis, Rapid repeat testing of grip strength for detection of faked hand weakness, *J Hand Surg [Br]* **25** (2000), 372–375.
[44] G. Waddell, J. McCulloch, E. Kummel and R. Venner, Nonorganic physical signs in low-back pain, *Spine*, **5** (1980), 117–125.
[45] G. Waddell, C. Main, E. Morris et al., Chornic low-back pain, psychologic distress, and Illness behavior, *Spine* **9** (1984), 237–241.
[46] G. Waddell and C.J. Main, Assessment of severity in low-back disorders, *Spine* **9** (1984), 204–208.
[47] A.P. Westbrook, M.W. Tredgett, T.R. Davis and J.A. Oni, The rapid exchange grip strength test and the detection of submaximal grip effort, *J Hand Surg [Am]* **27** (2002), 329–333.