Methodology Matters in Functional Capacity Evaluation

Determining Maximum Physical Capacity Through Scientific Methodology

Abstract

Functional Capacity Evaluations (FCEs) help physicians, employers, insurers, rehabilitation professionals, attorneys, and courts decide questions of return to work, work restrictions, disability, and rehabilitation planning. But every recommendation an FCE produces rests on one inference that measurement alone cannot supply: that the performance observed during testing actually represents the person's maximum physical capacity.

For four decades, FCE methodology has been an extended attempt to answer that question. Standardization made testing consistent but not necessarily accurate. Observation, heart rate, the coefficient of variation, and rapid exchange grip testing each tried to infer maximal effort from something other than the performance itself, and each was eventually shown to fall short: reproducible or reliable, perhaps, but not able to reliably tell maximal from submaximal effort. That pattern pushed the field toward a different approach — comparing independent measurements of the same physical ability against each other, rather than reading secondary signs believed to accompany effort.

That approach has since been tested directly. Studies of simultaneous bilateral hand strength testing and comparative lifting protocols found that independently generated measurements of the same physical capacity reliably agreed with each other during genuine maximal effort, and just as reliably diverged during instructed submaximal effort — evidence that this kind of cross-referencing can do what four decades of indirect indicators could not. The larger conclusion extends past any single test: the credibility of an FCE depends less on how many variables it measures than on whether the methodology connecting those measurements to a conclusion has itself been validated.

Introduction

Medicine depends on methodology. Whether clinicians are interpreting lab values, pulmonary function tests, imaging, or nerve conduction studies, they rely on methods whose accuracy has been established through research — not just on the numbers a test produces, but on confidence that the test measures what it claims to measure. Functional Capacity Evaluation deserves that same standard.

An FCE differs from most medical exams in one respect: it is not trying to diagnose disease. It is trying to determine what work a person can physically do, and to use that

determination to support decisions about return to work, permanent restrictions, disability, rehabilitation, and vocational planning — decisions that carry real medical, financial, and legal weight.

On its face, the task looks straightforward. The evaluator measures lifting, carrying, pushing, pulling, grip strength, and positional tolerance, then compares the results to the physical demands of a job. But the measurements alone do not answer the question that actually matters. A worker who lifts 50 pounds has, undeniably, lifted 50 pounds. What the test cannot tell you on its own is whether 50 pounds is that person's ceiling, or just what they chose to show that day.

That distinction sits underneath every conclusion an FCE produces. Work restrictions, disability opinions, vocational recommendations — all of them assume the measured performance reflects the person's actual physical capability. If it does not, everything built on top of it is weaker than it appears.

So the central problem in FCEs has never really been measuring performance. Reliable instruments for force, weight, distance, and time have existed for decades. The harder problem has been knowing whether what was measured represents the person's maximum.

Framed that way, the history of FCE methodology stops looking like a list of unrelated testing procedures and starts looking like a series of attempts to answer one question: how can an evaluator tell that observed performance represents maximum physical capacity?

Standardization tried to make testing procedures consistent. Observation tried to infer maximal effort from clinical judgment. Physiologic measures tried to correlate cardiovascular response with exertion. Statistical methods looked for reproducibility across repeated trials. Comparative methodologies, more recently, examined reproducibility across independent but equivalent tasks. Each was a different answer to the same underlying problem — some more useful than others, some reliable but not valid, some eventually shown to have real weaknesses despite wide clinical adoption.

This paper is not an argument for any particular commercial protocol, and it is not a criticism of the investigators whose work built the field. It is an account of how FCE methodology has moved, over time, from measuring physical performance toward validating the inferences drawn from those measurements — because ultimately, confidence in an FCE is confidence in its methodology, not just its numbers.

The Profession's First Response: Standardization

As FCEs developed during the 1980s, there was little consistency in how it was done. Testing sequences differed by facility, lifting procedures varied by evaluator, instructions were not standardized, and documentation varied widely. Two clinicians examining the

same person could easily reach different conclusions simply because they ran different tests.

The profession's first fix was standardization. Investigators such as Matheson[1] and Isernhagen[2] pushed for consistent instructions, testing sequences, progression of activities, termination criteria, and reporting — a common framework that made FCEs comparable across evaluators and facilities. This mattered enormously: without consistent administration, there is no way to know whether differences in outcome reflect differences in the patient or just differences in how the test was run.

But standardization solved only part of the problem. A test can be perfectly standardized and still be unable to tell you whether the performance it measured was someone's actual maximum. Consistency of administration is not the same thing as validity of inference — a distinction that shows up well beyond FCEs. A thermometer can report temperature to a tenth of a degree and a scale can weigh to the ounce, and both can be entirely reliable while still being the wrong tool for a different question. A standardized FCE can just as consistently record how much weight was lifted, how much force was applied, or how long a carry lasted — but that only describes what happened during the test. It does not establish, on its own, whether more was physically possible.

So standardization solved the problem of examiner variability without touching the deeper one: how does an evaluator know that what was observed was actually a maximum? The profession's next move was to reach for the tool every clinician already had — direct observation.

Observation as Evidence of Maximum Performance

Observation has always mattered in physical examination. Clinicians routinely read posture, gait, coordination, balance, movement quality, guarding, compensation, facial expression, and dozens of other cues that inform diagnosis and treatment — information instruments alone cannot capture. So it was natural that early FCE practice leaned heavily on clinical judgment: evaluators were expected to gauge maximal effort from body mechanics, movement speed, facial expression, muscle recruitment, breathing pattern, vocalization, apparent fatigue, willingness to continue, and similar behavioral signs. The underlying assumption was intuitive — an experienced clinician should be able to tell the difference between someone giving their all and someone not.

That intuition is not unreasonable on its face. People make observational judgments all the time. A mechanic hears an abnormal engine sound; a musician catches a slightly flat note; a physician often senses illness before the labs confirm it. Experience genuinely sharpens observation.

But the scientific question is not whether clinicians observe things — it is whether observation alone can reliably tell maximal from submaximal effort, and those are different questions. Recognizing poor lifting mechanics is not the same as knowing

whether more capacity exists. Noticing guarded movement is not the same as knowing whether the person is voluntarily holding back. Seeing apparent fatigue is not the same as confirming physiologic maximum has been reached. Each of those conclusions requires an inferential leap, not a direct reading.

That gap became a real problem as FCEs took on more weight in workers' compensation, disability determination, and litigation. Observations that once simply guided treatment planning were now supporting opinions about permanent restrictions, employability, and financial compensation. As the stakes rose, researchers started asking the obvious question: can trained observers actually tell, from observation alone, whether performance represents someone's true maximum?

Reliability Is Not Accuracy

Early studies of observational assessment tended to measure interrater reliability — how often different clinicians reached the same conclusion watching the same performance.[3] That seems, at first, like a reasonable measure of quality: if experienced evaluators consistently agree, the method should be trustworthy.

But agreement and accuracy are not the same thing. Reliability tells you whether observers agree with each other; it does not tell you whether they are right. Two clinicians can both conclude that someone gave maximal effort, and that agreement shows only that they are consistent with each other — not that either of them is correct. They could simply be making the same mistake in tandem.

This confusion between reliability and validity is not unique to FCE — it shows up throughout the history of science. Observers can reliably estimate an object's weight by eye while still being less accurate than a scale. Radiologists can reliably agree on a finding using criteria that later turn out to need revision. Agreement, on its own, does not establish truth.

Within FCE specifically, observational reliability turned out to be another real but incomplete advance: it improved consistency without answering the underlying question of whether measured performance was actually maximal.[4] The profession had gotten more consistent. It still had no objective way to tell whether what it measured was a ceiling or a choice.

That realization pushed the field to look past observation alone, toward methods that could test the evaluator's conclusion independently — physiologic response, behavioral indicators, statistical patterns, and reproducibility testing, each an attempt to answer the question observation by itself could not.

The Search for Objective Indicators of Maximum Performance

As observation alone became harder to defend scientifically, researchers looked for objective markers that might distinguish maximal from submaximal performance. The challenge was real: maximum physical capacity cannot be observed directly. It has to be inferred from something measurable that is presumed to accompany it — heart rate, statistical patterns in repeated trials, behavioral cues, patterns of force production. The logic was straightforward: if a measurable trait reliably differed between maximal and submaximal effort, it could serve as a proxy for maximum capacity.

The literature spent the next two decades testing that premise. Some approaches held up under limited conditions; others looked promising early on and fell apart under closer scrutiny. What they collectively established was a principle that turned out to matter more than any single test: a measurement can correlate with maximal performance without being able to classify it. That finding eventually pushed FCE research away from surrogate indicators and toward methods that could validate performance more directly.

Physiologic Measures

Heart rate was among the earliest candidates. The logic was intuitive — physical work requires energy, oxygen, and cardiovascular effort, so greater effort should mean a higher heart rate. It also had practical appeal: heart rate can be measured objectively and repeatedly, without the subjectivity of clinical observation.

The physiology, though, turned out to be messier than expected. Heart rate reflects far more than mechanical work — age, medication, anxiety, conditioning, hydration, temperature, pain, caffeine, nicotine, and underlying cardiovascular disease all move it independently of effort. Two people doing identical work can show very different heart rates. Even the same person can vary substantially from day to day under similar conditions.

Age-predicted maximum heart rate formulas made this worse, not better. They are useful for general exercise guidance, but their confidence intervals are wide enough — often several dozen beats per minute — that they cannot reliably separate maximal from submaximal effort in an individual patient. Later reviews concluded that heart rate, on its own, simply was not an accurate way to determine maximum effort during functional testing.[5] It remained useful for monitoring safety during testing. It did not solve the underlying problem.

Statistical Reproducibility: The Coefficient of Variation

A separate line of research looked at reproducibility across repeated measurements rather than cardiovascular response. The logic seemed sound: if someone were truly giving maximal effort, repeated trials should produce fairly consistent results; if effort varied intentionally, so should the numbers. That reasoning led to widespread use of the coefficient of variation (CV) — a statistic expressing how much repeated measurements

vary relative to their average — in grip strength and isometric testing. A low CV suggested consistent, presumably maximal, performance; a high CV suggested the opposite. It felt objective in a way clinical observation did not.

The assumption did not hold up. Repeated studies found that people giving deliberately submaximal effort could reproduce that submaximal performance with remarkable consistency, while some people genuinely giving maximal effort showed more variability than expected — driven by pain, unfamiliarity with testing, fatigue, and ordinary biologic variation, independent of effort itself.[6,7] In other words, a low CV did not prove maximal effort, because consistent submaximal performance is still statistically consistent. And a high CV did not prove submaximal effort, because normal variability can look the same way.

Lechner and colleagues reviewed the evidence and concluded that the CV simply was not a reliable way to determine sincerity of effort[5] — a conclusion later investigations reinforced.[6,7] The broader lesson extends past the CV itself: reproducibility alone does not establish validity. Data can be perfectly internally consistent and still fail to answer the question being asked of it.

Rapid Exchange Grip Testing

Rapid Exchange Grip (REG) testing took a different approach. Instead of looking at consistency across repeated maximal efforts, it compared force during conventional grip testing against force generated during fast, alternating contractions between the hands, on the theory that explosive alternating grips should produce less force than sustained maximal ones. Unexpectedly high force during rapid exchange was taken as a sign of inconsistent effort.[8] Like the CV, it had intuitive appeal and caught on quickly.

Closer study found real problems. The test was hard to standardize — variations in speed, rhythm, timing, and how the hand interacted with the dynamometer all introduced noise, and mechanical impact against the instrument could produce readings unrelated to actual grip force. More fundamentally, multiple studies found the test lacked the sensitivity and specificity to reliably separate maximal from submaximal effort. As with the CV, the issue was not measurement error or bad math — it was that the variable itself was not discriminating enough to support the conclusion being drawn from it.

Lechner's 1998 review advised clinicians against relying on rapid exchange grip testing, the coefficient of variation, or bell-curve analysis[9] to judge sincerity of effort, since the evidence did not support any of them for that purpose.[5] That conclusion mattered beyond the specific tests it named. It suggested the field was not just picking the wrong tests — it might be pursuing the wrong strategy altogether. Attention began shifting away from indirect indicators of effort and toward a different question: could maximum physical capacity be validated by comparing independent measurements of the same physical ability, rather than by reading secondary signs believed to accompany effort? That question would reshape the direction of FCE research.

A Different Scientific Approach

By the late 1990s, a pattern had become clear. Observation, physiologic measures, and statistical analyses were all trying to infer maximum effort from variables that merely accompanied performance, rather than from performance itself. The evaluator observed movement quality, measured heart rate, calculated variability, or read force patterns — then inferred whether what was measured represented a true maximum.

The problem was not that these measurements were inaccurate. Heart rate can be measured precisely; statistical variability can be calculated exactly; observations can be recorded consistently. The problem was the inference sitting on top of them. None of these measurements directly established that someone had reached maximum capacity — each depended on an assumption about how the measured variable related to the thing actually being asked about.

That left the field with a choice: keep trying to refine indirect indicators, or question whether indirect indicators were the wrong tool altogether. Researchers began exploring the second option — asking not whether secondary signs suggested maximal effort, but whether the physical performance itself could be verified independently.

Independent Comparison as Scientific Validation

Independent confirmation is a familiar principle in science. A lab result gets checked by repeat analysis on a different instrument. Survey measurements get confirmed from multiple reference points. Engineers cross-check calculations that should converge on the same answer. Confidence rises when different methods measuring the same thing agree.

Medicine works the same way. Radiographic findings are read alongside the physical exam. ECG abnormalities are weighed against clinical presentation. Pulmonary function tests are interpreted together with symptoms and imaging. Independent observations that converge strengthen the conclusion.

FCEs, historically, did not have this. Most traditional methods looked at performance from a single angle — a worker lifted a crate, grip was measured on one side, isometric force was recorded once — and the evaluator then tried to judge whether that isolated number represented a true maximum.

The next development broke from that pattern. Instead of scrutinizing a single measurement more closely, researchers asked whether the same physical ability could be measured twice, under independent conditions, and then compared. If both measurements genuinely reflected maximum capacity, they should agree within a predictable range. If they did not, the disagreement itself became the finding worth investigating. That is a real conceptual shift: the method no longer depends on reading secondary signs of effort. It

depends on whether two independent measurements of the same ability reproduce each other.

The Concept of Distraction-Based Comparison

Distraction testing has deep roots in clinical medicine. Its best-known application is probably Waddell and colleagues' work on low back pain, which introduced distraction testing through the seated straight-leg raise, or “flip test” — comparing a patient's response under two different testing conditions. If a finding improved under distraction, that suggested the original result should be interpreted with caution, not because the patient had been caught lying, but because two independently generated observations disagreed.[10]

It is worth being precise here: Waddell never proposed his nonorganic signs as tests of sincerity or malingering. Their purpose was to flag patients who might benefit from psychological assessment, and to account for the role nonorganic factors can play in a clinical exam.[10] Using them later as a direct measure of effort stretched them well past what they were designed to do — a point Lechner and colleagues make explicitly in their review.[5]

What Waddell's work really contributed, methodologically, was the idea of independent comparison: rather than relying on one observation, compare two exams that should produce similar findings if the underlying condition has not changed. Agreement builds confidence; disagreement prompts a closer look. Though it originated in spinal examination, that principle pointed toward a different way of approaching FCEs altogether. Instead of asking whether a single lifting task looked maximal, why not compare two independent lifting tasks that draw on the same underlying capacity? Instead of inferring effort from heart rate or facial expression, why not check whether two independent measurements of the same ability actually match? The focus shifted from observing effort to testing reproducibility — a shift that became the foundation for the next generation of FCE research.

Application to Hand Strength Assessment

The hand turned out to be an ideal place to test this idea. Traditional grip testing relied on one-sided measurements read through bell-curve analysis, coefficient of variation, rapid exchange grip testing, or some combination — and, as the literature increasingly showed, these approaches often failed to separate maximal from deliberately submaximal effort.[5] Grip strength is also easy to quantify and control, and hand biomechanics allow several ways of measuring essentially the same underlying ability — a good combination for testing a new idea experimentally.

So researchers asked a different question: would the maximal force produced during conventional one-sided testing reproduce during simultaneous bilateral testing, if both were measuring the same underlying capability? Unlike repeating the same unilateral

trial, simultaneous bilateral testing changes the testing condition while measuring the same physiologic capacity — the subject is not just trying to repeat an identical movement, but generating two genuinely independent measurements that can then be compared. That is the key difference: the method no longer leans on assumptions about variability, heart rate, or examiner judgment. It leans on whether two independent expressions of the same physical capability agree.

The validation study found that independently derived grip measurements consistently reproduced within defined limits during maximal effort, while instructed submaximal efforts produced a characteristic pattern of disagreement — correctly sorting instructed maximal from instructed submaximal effort with exceptionally high classification accuracy.[11] No method is perfect in every circumstance, but the results discriminated far better than the surrogate indicators that preceded it. That is what validity means in practice: whether a method actually measures what it claims to measure. By that standard, the bilateral hand strength protocol had answered the question this paper opened with — under the tested conditions, independently derived measurements correctly identified true maximal and true submaximal effort. Reliability — whether different examiners administering the same protocol arrive at comparable results — is a related but separate question, and this paper does not claim it was resolved here. The two are not interchangeable: a method can be valid but poorly standardized, or well standardized but not valid. What the hand strength studies established was the harder of the two: validity itself.

The hand studies filled a gap earlier FCE research had not closed: a method that demonstrated it could classify performance accurately, using independently generated measurements rather than an examiner's impression. Whether the protocol is equally reliable across different examiners is a fair question for future research to confirm — but it is a distinct question from the one these studies actually answered.

That question was tested directly. A follow-up study applied the same validity criteria to 200 consecutive clients undergoing an FCE in actual clinical practice — not instructed volunteers, but real claimants with real injuries. Clients who failed the criteria did report more pain and more visible pain behavior, which cuts either way. But they also had a lower rate of documented surgical history than clients who passed — the opposite of what you'd expect if failing the criteria simply reflected genuine, pain-limited performance rather than effort regulation.[12] The pattern held outside the controlled setting where the method was first validated.

Extending the Principle Beyond the Hand

The hand studies were compelling, but an obvious question remained: would the same principle hold for larger, more complex activities? Grip strength is only one piece of occupational performance — most work-related disability decisions hinge far more on lifting, carrying, pushing, pulling, and whole-body material handling. If independent comparison was a genuine methodological advance, it needed to work beyond dynamometry.

Dynamic lifting is a much harder test of the idea. Multiple joints contribute to force production at once; balance, coordination, strategy, body size, limb length, and movement speed all shape performance; fatigue sets in faster; and testing has to account for progressive loading while keeping the patient safe. Simply repeating the same lift would not be enough — repetition invites practice effects and conscious recall of prior attempts, which is not genuine independent confirmation. Instead, researchers looked for lifting tasks that differed operationally but drew on the same underlying physical capacity, so a subject could not just mechanically repeat a previous effort. Agreement between the two, under those conditions, would mean something. As with the hand studies, the goal was not to observe effort — it was to determine whether two independently measured expressions of lifting capacity confirmed each other.

Dynamic Lifting Validation

The original lifting investigation found that independently derived lifting measurements reproduced predictably during maximal performance, while instructed submaximal efforts produced characteristic disagreement between the comparison tasks.[13] It correctly identified every instructed-maximal participant in the study while also distinguishing instructed submaximal performance. What mattered beyond the classification numbers themselves was that the method no longer required the evaluator to read secondary signs of effort — the conclusion came from the objective relationship between two independently measured lifting performances, not from judgments about whether someone looked fatigued or motivated.

Later studies pushed on the robustness of the approach rather than treating the original results as the final word — testing whether it held up across different populations, testing environments, and examiner groups. That is the normal path scientific evidence takes: establish feasibility first, then test reproducibility, then explore how far the finding generalizes. The dynamic lifting literature followed exactly that path.

Continued Investigation of Comparative Lifting Methodology

Confidence in a methodology does not come from one study — it builds as later investigations keep supporting the original finding under more varied, more realistic conditions. That is what happened with comparative lifting. After the original validation study, researchers moved past tightly controlled laboratory conditions to test the method across a wider range of loading strategies, decision rules, and instructed performance levels, asking whether it would keep classifying performance accurately as testing got closer to real clinical practice.

The evidence said yes. One study examined progressive loading and found that people giving maximal effort kept showing the expected agreement between independent lifting measures even as physical demand increased[14] — the method held up as loading progressed, supporting the idea that the agreement reflects real physical capacity rather

than something specific to one lifting task. A second study focused on decision-making criteria,[15] developing explicit classification rules for interpreting agreement and disagreement between independent tasks. That mattered because objective measurement alone does not remove subjectivity if interpretation stays inconsistent — standardized decision criteria make sure equivalent findings lead to equivalent conclusions no matter who is evaluating. Perhaps the most clinically relevant study tested whether the method could distinguish instructed submaximal performance during actual dynamic lifting,[15] and found the same pattern as the original work: maximal effort reproduced predictably across independent tasks, while instructed submaximal effort produced the same characteristic disagreement.

Together, these studies reinforced the underlying principle by showing it held up despite changes in testing conditions and analytical approach — proof of concept, then reproducibility, then generalizability, which is how confidence in a scientific method is supposed to build.

The Scientific Significance of Independent Cross-Reference

What comparative hand and lifting methodologies contributed is best understood as a change in reasoning, not just a new set of tests. Earlier methods looked for evidence that maximal effort had occurred. Comparative methods look for evidence that measured physical capacity reproduces independently — and that is a real difference. When heart rate, facial expression, or statistical variability are used to judge performance, the conclusion depends on an assumption about how that variable relates to maximum effort. The assumption may be reasonable, but it is still indirect. Independent comparison removes much of that inferential chain: instead of asking whether a secondary sign suggests maximal effort, it asks whether two independent measurements of the same physical capability confirm each other. The evidence comes from the performance itself, not from something presumed to accompany it.

This mirrors a principle found throughout empirical science: independent observations that converge on the same conclusion generally carry more weight than repeated interpretation of a single observation. The value is not just repetition — it is the independence of what is being compared. Agreement reached under different testing conditions makes it less likely the result is chance, examiner bias, or an artifact of one particular procedure.

Within FCEs, this amounts to a shift from interpretive methodology toward comparative methodology, and it changes the evaluator's role. Rather than judging whether someone appears to be giving maximal effort, the evaluator checks whether independently generated measurements agree to the degree expected if maximum capacity was actually expressed. The conclusion depends less on examiner opinion and more on a measurable relationship between independently obtained data — which matters in a field where the stakes are often financial, vocational, or legal. No methodology eliminates every source

of uncertainty, and none should be expected to; human performance is genuinely variable, and no method will classify every person correctly in every circumstance. But a method does not have to be perfect to represent real progress. Its job is to reduce uncertainty by replacing unsupported inference with evidence that can actually be tested — and that is what comparative methodology does: it shifts the emphasis from interpreting isolated observations to evaluating whether independently measured physical capacity reproduces itself. That is a methodological advance, not just a procedural one.

Conclusion

The implications are not only scientific. FCEs shape decisions about return to work, permanent restrictions, rehabilitation, disability status, and, often, litigation — decisions that affect employment, income, and long-term quality of life. That does not mean every FCE has to follow identical procedures; different protocols reasonably emphasize different goals depending on the referral question and setting. But it does mean evaluators should understand the assumptions built into whatever method they are using. Observation assumes a clinician can accurately recognize maximal effort. Physiologic monitoring assumes cardiovascular response reflects physical capacity closely enough to be useful. Statistical reproducibility assumes consistency distinguishes maximal from submaximal performance. Comparative methodology assumes that independent measurements of the same physical capability should agree within predictable limits when a true maximum has been reached. Every one of those assumptions is testable — and that is the real point of this review. Two evaluators can take similar measurements and still reach different conclusions if they are working from different analytical frameworks, which means disagreement between FCEs often reflects a difference in methodology, not a difference in the patient's actual performance. So the more useful question usually is not which protocol is superior, but whether the assumptions behind a given method are actually supported by evidence.

Comparative methodology answers that problem more directly than anything that came before it — not because it measures something new, but because it tests whether what was measured can be trusted. That is the standard the field should keep moving toward. The next real advance in Functional Capacity Evaluation is unlikely to come from adding another variable to measure. It will come from continuing to demand that the reasoning connecting a measurement to a conclusion can itself withstand scrutiny.

References

1. Matheson LN, Ogden LD. Work Tolerance Screening. Trabuco Canyon, CA: Rehabilitation Institute of Southern California; 1983.

2. Isernhagen SJ. Functional capacity evaluation: rationale, procedure, utility of the kinesiophysical approach. J Occup Rehabil. 1992;2(3):157-168.

3. Isernhagen SJ, Hart DL, Matheson LN. Reliability of independent observer judgments of level of lift effort. Work. 1999;12(2):145-150.

4. Schapmire DW, St. James JD, Townsend R, Feeler L. Accuracy of visual estimation of effort in classifying a lifting task. Work. 2011;40(4):445-457.

5. Lechner DE, Bradbury SF, Bradley LA. Detecting sincerity of effort: a summary of methods and approaches. Phys Ther. 1998;78(8):867-888.

6. Niebuhr BR, Marion R. Detecting sincerity of effort when measuring grip strength. Am J Phys Med Rehabil. 1987;66(1):16-24.

7. Townsend R, Schapmire DW, St. James J, Feeler L. Isometric strength assessment, Part II: static testing does not accurately classify validity of effort. Work. 2010;37(4):387-394.

8. Hildreth DH, Breidenbach WC, Lister GD, Hodges AD. Detection of submaximal effort by use of the rapid exchange grip. J Hand Surg Am. 1989;14(4):742-745.

9. Stokes HM. The seriously uninjured hand--weakness of grip. J Occup Med. 1983;25(9):683-684.

10. Waddell G, McCulloch JA, Kummel E, Venner RM. Nonorganic physical signs in low-back pain. Spine (Phila Pa 1976). 1980;5(2):117-125.

11. Schapmire DW, St. James JD, Townsend R, Stewart T, et al. Simultaneous bilateral testing: validation of a new protocol to detect insincere effort during grip and pinch strength testing. J Hand Ther. 2002;15(3):242-250.

12. Schapmire DW, St. James JD, Feeler L, Kleinkort J. Simultaneous bilateral hand strength testing in a client population, Part I: diagnostic, observational and subjective complaint correlates to consistency of effort. Work. 2010;37(4):309-320.

13. St. James JD, Schapmire DW, Feeler L, Kleinkort J. Simultaneous bilateral hand strength testing in a client population, Part II: relationship to a distraction-based lifting evaluation. Work. 2010;37(4):395-403.

14. Swift MC, Townsend R, Edwards DW, Loudon JK. Decision-making data: expectations for reproducibility of lifting on separate days. Prof Case Manag. 2018;23(4):204-212.

15. Townsend R, Bell S, Harry J. Accuracy of distraction based lifting criteria for the identification of insincere effort utilizing the under loading method. Work. 2016;55(4):873-882.