VCU

Virginia Commonwealth University

Functional Capacity Evaluation Course

Williamsburg, Virginia
September 14, 2000

ARCON
VerNova FCE

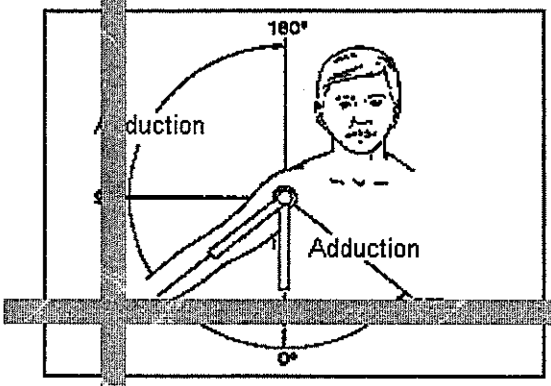
Part 2 Scientific Foundations

scientific foundations

I. Measurement
II. Reliability
III. Validity
IV. Standardization

WHAT YOU NEED TO KNOW

- definitions
- practical examples
- how to maintain scientific principles in your practice



I. MEASUREMENT

MEASUREMENT SCIENCE

Measurement is the process of assigning numbers to the properties of objects, organisms or events. It must be quantifiable so a meaningful interpretation can be given between any two scores. The measurement system must include *quantification*, *objectivity, economy, communication* and *scientific generalization.*

Evaluation is the process of making judgments about the results of measurement. Hence a Functional Capacity *Evaluation* without measurement is not an evaluation, but an observation and description of evaluee behavior. In the absence of valid and reliable measures, the results of an assessment have greatly diminished significance.

DEFINITIONS

Measurement is the process of assigning numbers to the properties of objects, organisms or events. It must be *quantifiable* so a meaningful interpretation can be given between any two scores. The measurement system must include *objectivity*, *quantification*, *communication*, *economy* and *scientific generalization.*

Evaluation is the process of making judgments about the results of measurement.

Statistics are the communication system used to convey meaningful interpretation of the measurement

Standardization is the consistent application of test protocols to limit reliability error.

Reliability is the consistency of scores of an evaluee's performance on a test.

Construct Validity is the degree to which a test measures traits that are a theoretical representation of what is purported to be measured. Constructs cannot be observed, therefore it is necessary to measure behaviors that are believed to depict the construct.

Content Validity is the degree to which some sample of tasks are representative of content domain of what is being studied. Content validity focuses on test *forms* rather than test *scores*, and *instruments* rather than *measurements.*

Criterion validity is the extent to which *scores* on a test are related to a criterion measure.

Norm-Referenced Measurement is a process of defining a sample of a population, analysing the measurement of that sample, establishing the *generalizability* of that sample to the population and to other samples, and developing statistics to represent that sample and compare an individual score against that norm-reference.

Criterion-Referenced Measurement yields measurement that is directly interpretable in terms of specified performance standards. The measurement is independent of other evaluee scores whereas norm-referenced measurement is dependent on other evaluee scores.

QUANTIFICATION

Quantification permits greater levels of precision for reporting results. Powerful methods of mathematical analysis can be applied when the results are quantified.

Isometric strength can be measured and force quantified (lb). Dynamic strength factors of weight and distance can be measured to calculate biomechanical joint forces (ft-lb) and metabolic demand (kcal).

OBJECTIVITY

Measurement leads to objectivity by allowing reproduction of results for verification.

Strength measurement can be duplicated by two evaluators when measurement systems are utilized.
Force measurement will be consistent while qualitative rankings (strong to weak on a 5 point scale) will have greater variance.

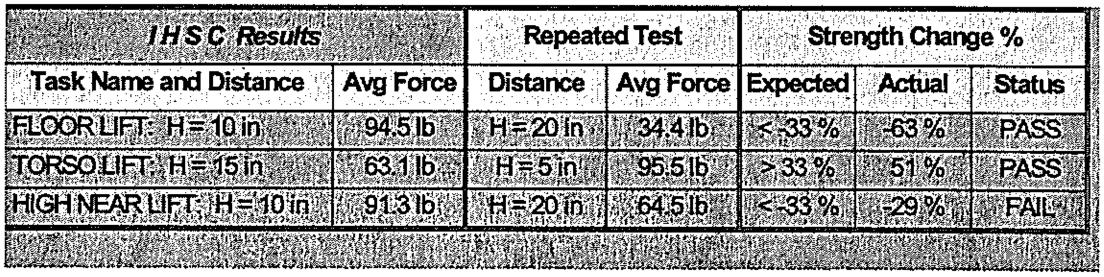
ECONOMY

Functional capacity evaluation faces shrinking reimbursement rates. The evaluator's challenge is to spend the minimal amount of time necessary to accurately predict an evaluee's ability to return to occupational tasks. Fewer measurements are required in each evaluation when measurements can be compared across a population sample. Predictive measures are more economical of time than subjective measurement.

Strength measurement can be taken at a single reference and extrapolated to the wealth of comparative measurement data.

COMMUNICATION

Statistics is a method of summarizing and analyzing data for purposes of interpretation. Statistics used to describe a set of test scores are referred to as descriptive statistics. Inferential statistics are used when test scores are used to make an inference based on the information from the evaluation. Statistics are the communication system used to convey meaningful interpretation of evaluation measurement.



I H S C Results		Repeated Test		Strength Change %		
Task Name and Distance	Avg Force	Distance	Avg Force	Expected	Actual	Status
FLOOR LIFT: H = 10 in	94.5 lb	H = 20 in	34.4 lb	< -33 %	-63 %	PASS
TORSO LIFT: H = 15 in	63.1 lb	H = 5 in	95.5 lb	> 33 %	51 %	PASS
HIGH NEAR LIFT: H = 10 in	91.3 lb	H = 20 in	64.5 lb	< -33 %	-29 %	FAIL

FUTURE DIRECTIONS IN FCE MEASUREMENT

Theoretical measurement research has led to many statistical analyses that apply to functional capacity evaluation test development. Reliability coefficient, standard error of measurement, analysis of variance, generalizability theory, rasch rating analysis, item response theory, repeated measures testing and sequential probability ratio test all hold promise to increase the scientific foundations of the FCE.

SCIENTIFIC GENERALIZATION

Any well developed test is based on a measurement model. Test validity and reliability dictated by the model establishes the procedures for estimating these characteristics for a specific evaluee.

Strength measurement research was undertaken in lifting planes that must be replicated in the evaluation to assure validity.

CLASSICAL TEST THEORY

Most tests used in functional capacity evaluation are based on Classical Test theory, the basis of test measurement for over 75 years. Classical Test theory is built around the concept of a true score and an error score. The true score represents the score the evaluee would have made if there was no measurement error. The error score is attributed to the measurement error.

Measurement models can be classified into two categories: *weak true-score* models and *strong true-score* models. The assumptions behind a weak model are not rigorous and can be met by many conditions, while the assumptions of a strong model are satisfied only under limited conditions. An example of a weak true-score model is a clinical/behavioral observation of a functional task such as walking. A standardized static lift protocol scored against a criterion dataset is an example of a strong true-score model.

Strong True Score	**Weak True Score**
Timed against criterion	**Observation only**
Standardized protocol	**Subjective**
Automated (computerized)	**Extemporaneous Record**

scientific foundations

II. RELIABILITY

MEAUREMENT MUST BE RELIABLE

If a test or measuring instrument is reliable it will consistently provide the same measurement of an evaluee. This is important because the score obtained is supposed to be a good indicator of an evaluee's true ability. Without reliability a test cannot measure performance against any standard. A hand grip dynamometer that is not calibrated, and has electronic interference contaminating the results, will give scores with a high error. Comparison of those scores to an established criterion will give false conclusions.

RELIABILITY INFLUENCES VALIDITY

Reliability influences validity. **For a test to be valid it must be reliable, however a reliable test is not necessarily valid.** A hand grip test can be performed reliably with proper calibration and setup, yet it will never be a *valid* measure of walking capability.

TYPES OF RELIABILITY

There are different measures of reliability. *Test-retest reliability* is the relationship between repeated measures. If the test gives statistically similar scores each time a *stable* function is tested then the test would have a strong coefficient of stability. However, many evaluees assessed in the FCE process are not medically stable, or do not apply consistent effort. Hence, it has been hypothesized that a test with inherently strong stability coefficient of variance performed on a medically stable, unimpaired biomechanical function, can be an indicator of effort reliability.

Inter-rater reliability is the coefficient of variance between two evaluators measuring the same evaluee. Evaluator bias is a threat to inter-rater reliability and needs to be controlled by use of objective versus subjective measurement protocols. Variance is diminished by standardization, training and automation of measurement. Criterion rating also diminishes evaluator bias, as the evaluators use equivalent benchmarks to measure the evaluee.

Equivalent form reliability is the coefficient of equivalency of two alternate forms of the same test. Although this methodology has not been utilized for Functional Capacity Evaluation, it holds promise for further measurement of evaluee consistency.

III. VALIDITY

WITHOUT VALIDITY THE EVALUATOR CANNOT MAKE INFERENCES

Validity is the most important quality of a test. It determines if the evaluator is measuring what is intended to be measured. The need for validity arises from the fact that an evaluee's functional ability at their specific continuous work tasks is not able to be directly measured. The evaluator is not afforded the time or opportunity to perform an assessment in this manner. The *Standards for Educational and Psychological Tests*, published by the American Psychological Association (the standard to which most human performance testing is held, including employment and exercise science testing) defines validity as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences."

TYPES OF VALIDITY

Construct validity is the process of determining the degree to which the test measures the construct it was designed to measure. Construct evidence is often gathered by also demonstrating that there is no relationship between the construct and theoretically divergent measures. An example of the divergent hypothesis is that strength (measured by an isometric load cell) is a construct inherent in lifting ability, but that it would have no relationship to intelligence.

Content validity is evidence that the test samples tasks that represent the domain being evaluated. The focus is on the instruments rather than the measurements. Content validity is important but does not in itself ensure validity. Indeed, the term '*face validity*' was concluded to be unacceptable as a basis for interpreting scores as early as the 1974 *Standards for Educational and Psychological Tests* published by the American Psychological Association. C*ontent validity* was dropped in favor of *content-related evidence* in the 1985 edition. An example of content validity in functional capacity evaluation is isometric strength testing versus dynamic strength testing. A dynamic strength test that involves lifting a weighted container has greater content relevance to lifting in the workplace than isometric test that involves pulling on a strain gauge. However, the score from a dynamic strength test still must be demonstrated to be representative of actual lifting ability in the real world of work via criterion related evidence.

TYPES OF VALIDITY

Criterion-related evidence quantifies the evidence that test scores will relate highly between tests measuring the same characteristic. For example an evaluee who performs well on a treadmill test measuring Vo$_2$ max would be expected to perform well on a carrying task requiring aerobic endurance. The evidence for criterion validity is gathered via *concurrent* validity designs, such as that described above, or by *predictive* validity designs, such as comparison of test scores to an established criterion of acceptable performance. The predictive validity design for the treadmill test would be to compare scores to a gas-exchange measurement.

Criterion-referenced tests are constructed that measurements are directly interpretable in terms of specific performance standards. Representative samples of the tasks are organized into a test. Measurements taken are used to make a statement about the performance of the evaluee relative to that domain. The domain criterion can be used as a cut-off score to make decisions concerning evaluee capacity to perform the task relative to the criterion. However in a rehabilitation approach the criterion can suggest not only how divergent the evaluee is from satisfactory performance, but also what accommodations, modifications and engineering aides might be useful.

Norm referenced validity is the process of defining a sample of a population, analyzing the measurement of that sample, establishing the *generalizability* of that sample to the population and to other samples, and developing statistics to represent that sample and compare an individual score against that norm-reference. For example, a handgrip test score can be described at a 50th percentile compared to a norm group. This is useful in establishing the general ability the evaluee has, but it is not specific to a work criterion measure.

The Construct of strength is comprised of measureable Content (e.g. force, power, metabolic endurance, joint stability and range of movement). Each Content domain has been researched to establish Criterion and/or Norm referenced evidence.

Protocols used in the research form the Standardization for Reliability of evaluation measurement.

IV. STANDARDIZATION

CONSISTENCY LIMITS RELIABILITY ERROR

Standardization is the consistent application of test protocols to limit reliability error. The first step of standardization is a set of instructions for test administration that needs to be followed precisely each time the test is administered. The clinical environment, materials and equipment should remain the same from one administration to the next. Scoring should be performed with a predetermined protocol as well.

STANDARDIZATION COMPONENTS

Standardization involves the establishment of scores against which to compare evaluee scores.

Standardization should diminish evaluator bias. Judgment on evaluee score should rest primarily with the rating system, rather than on subjective opinion of the evaluator. When evaluator subjective opinion and evaluee behavior rating are collected, a standardized checklist format is preferable.

Methods of achieving standardization have been demonstrated via computer automated data collection from electronic measuring devices. Studies have shown that bias reliability is increased, *if training is sufficient*. Training can be standardized by requirements for general clinical and specialized training certification. Manufacturer training for test devices is also necessary. Training materials and multi-media format assist in standardization.

When an atypical evaluee presents for assessment, measures should be taken to control for threats to standardization. A common example is to request a trained medical interpreter for testing an evaluee who does not have sufficient language skills to communicate with the evaluator. If test modification is necessary for an atypical evaluation (eg. a one handed carry test for an amputee) then results can only be interpreted with usual predictive power if the test modifications can be applied within the test protocol. Otherwise a "reader caution" must be applied in the report *(I.e. "This test was modified to accommodate the evaluee's disability. True ability may not be accurately represented by this measurement.")*.

SUMMARY

The learning objective of this section was to:

- ✓ Introduce measurement concepts
- ✓ Acquaint the evaluator with validity, reliability and standardization
- ✓ Outline the major measurement issues in the FCE

LEARNING EXERCISE:

Design a novel test to measure sobriety. Consider the following features the test must have:

Construct Validity: What assumption do you make between the measurement and the level of sobriety/drunkenness?

Content Validity: What items or instruments in the test represent the condition of sobriety versus drunkenness?

Would you use a *norm* or *criterion* to compare the test scores to?

If you develop a norm would it be *valid* to use a rugby club after two hours of post-game celebration in a pub or a church choir during a Wednesday night practice?

If you use a criterion, indicate how you would develop the criterion and score the test against the criterion.

What criterion is commonly used to pass/fail sobriety in your state/province/country?

Reliability: Design a measurement device using available materials that will demonstrate consistent scores from test to retest and between evaluators.

Standardization: Describe the standard protocol to be followed.

scientific foundations

REFERENCES

1. Amer. Educational Research Assoc., Amer. Psychological Assoc. & National council on Measurement in Education. (1985). "Standards for educational and psychological testing". Washington, D.C.: Authors.
2. American Psychological Association. (1985) *Standards for educational and psychological tests*. Washington D.C.
3. Botterbusch, Karl.and Michael, Nancy. (1985) *Testing and Test Modification in Vocational Evaluation*. Materials Development Center, University of Wisconsin-Stout, Menomonie, WI
4. Cronbach, L.J.(1971). "Test validation". In R.L. Thorndike (Ed). Educational Measurement (2nd Ed.) Washington, D.C.: Amer. Council on Education.
5. Cronchback L, et al. Dependability of behavioral measurements: theory of generalization for scores and profiles. (New York: John Wiley and Sons, 1972).
6. Fisher WP, jr. "Objectivity in measurement: A philosophical history of Rasch's separability theorem." M. Wilson, ed. Objective measurement: theory into practice. (Norwood, N.J.: Ablex, 1992), 29-55.
7. Kraiger, Kurt; Teachout, Mark S; "Generalizability Theory as Construct-Related evidence of The Validity of Job Performance Ratings". Human -Performance. 1990; Vol 3 (1): 19-35.
8. McHenry, Jeffrey J. et-al. "Project A Validity Results: The Relationship Between Predictor and Criterion Domains". Personnel-Psychology. 1990 Sum; Vol 43(2): 335-354.
9. Messick S. "The standard problem: Meaning and values in measurement and evaluation." Am Psychol 30 (1975): 995-966.
10. Messick, S. (1995). "Validity of psychological assessment: Validation of inferences from person's responses and performance as scientific inquiry into scoring meaning". American Psychologist, 9, 741-749.
11. Moss, P.A. (1994). "Can there by validity without reliability?". Educational Researcher, 23, 5-12.
12. Ritchie, Richard J; " Using The Assessment Center Method to Predict Senior Management Potential", Consulting Psychology Journal: Practice and Research. 1994 Win; Vol 46(1): 16-23.
13. Safrit, M.J., and Wood, T. M. (ed.) (1989). *Measurement Concepts in Physical Education and Exercise Science*. Champaign, IL: Human Kinetics Books
14. Tomarken, Andrew J; "A Psychometric Perspective on Psychophysiological Measures", Psychological-Assessment. 1995 Sep; Vol 7(3):387-395
15. Walter Pruit. (1986) *Vocational Evaluation*. Walt Pruit Associates. Menomonie, WI
16. Wise, Lauress L; McHenry, Jeffrey J; Campbell, John P. "Identifying Optimal Predictor Composites and Testing For Generalizability Across Jobs and Performance Factors". Personnel-Psychology. 1990 Sum; Vol 43(2); 355-366.