LETTERS TO THE EDITOR

Letters to the Editor: Rehabilitation Idea Exchange

A publicly available record of scientific and engineering results is vital to rehabilitation progress and represents the fundamental commitment of the *Journal of Rehabilitation Research and Development*. Ideas, intellectual debate, and other information also are important to progress. Thus, to give our readers an opportunity to present new ideas, alternative points of view on existing ideas, and commentary on material published, we are initiating a "Letters to the Editor" department to serve as a "Rehabilitation Idea Exchange."

Letters should be typewritten double-spaced. Possible conflicts of interest should be described.

Only original material not published elsewhere will be accepted.

Letters will be refereed and indexed as deemed appropriate.

Re: Qualitative and Quantitative Gait Phase Analysis by Continuous Monitoring of Inter-Ankle Distance*

To the Editor:

Pinzur et al. are to be congratulated in developing a relatively simple, inexpensive "temporal recorder of gait." The system appears to be reliable and easy to use. For these reasons I feel that it could have wide applicability to a variety of research and clinical investigations related to both normal and abnormal gait. However, in attempting to conceive of its applicability, I have formulated some questions that I hope the authors can answer.

First, Pinzur et al. state that "the time required for the acoustic energy to propagate from transmitter transducer to the receiver transducer is a measure of the distance between two transducers." From measurements so obtained and plotted as in Figure 4, it is easy for me to understand from Figure 1 how different temporal events for normal gait can be determined. However, it is less clear how forward progressive measurements can be calculated to determine forward velocity (see RESULTS p. 52) if only this device is utilized. The distance transversed by the acoustic wave is not parallel to the forward direction; thus a component of lateral distance must always exist in the resultant distance this device measures. One could therefore conceive of a situation in which this device could calculate, the same forward velocity for two people walking with the same cadence but where one is walking with a shorter forward stride length and a longer stride width than the other; in reality these two people have two different velocities.

Second, if one compares Figure 4 with Figure 5 it appears that the former is an idealized model of the way the data should look, but in reality the latter is a better representation of actual recordings. Is this true? If it is, I would have a great deal of difficulty discerning the true time of change in slope of the IAO upward and/or downward tracings with which to define heel-strikes and toe-offs. I am particularly confused by Figure 5; because according to how such information is verified in this study, results in when EMG activity in the anterior tibialis is active. In Figure 5, "the block marks stance phase," yet the "filtered EMG original" seems to show activity of this muscle only during weight acceptance (one of the two double-limb stance phases) and not during any swing phase when this muscle in normal subjects is almost always also active. The lower channel demonstrating "electrical activity recorded from a surface electrode mounted at the motor point....in the tibialis anterior," is also of little help as it appears that the filtered EMG activity shown above corresponds to times of low frequency–high amptitude signals. Could this be motion artifact?

Finally, even the block marks themselves do not end at times when the downward slope of the IAO changes. If this is a typical tracing from a normal subject, I cannot help but wonder how difficult it might be to interrupt such curves from subjects with pathlogical gaits.

* PINZUR MS, LEVINE PD, TRIMBLE J, HAGG K, SHERMAN R: Qualitative and quantitative gait phase analysis by continuous monitoring of inter-ankle distance. *J Rehabil Res Dev* 21(2): 50–53, 1984.

Journal of Rehabilitation Research and Development Vol. 22 No. 3 July 1985

Although other more technical questions (probably minor ones) also arise, the two aspects I have just cited raise grave doubts about the usefulness of this device. I hope I am incorrect since this is certainly an appealing device, and I strongly hope the authors will allay my concerns in their reply.

Sheldon R. Simon, Ph.D
The Children's Hospital Medical Center
300 Longwood Avenue
Boston, Massachusetts 02155
Gait Analysis Laboratory
at 52 Binway St.

The Fallacy of Timed Functional Tests

To the Editor:

These comments are presented in response to a persistent notion among clinical researchers that performance testing must be based on measurement of performance speed. The difficult task of measuring performance is examined in the context of the more general problem of developing a mathematical model for a human operator.

Drawing on experiences in the mathematical modeling of human operators and on clinical experience in performance testing, we call attention to the inaccessibility of direct evidence of intent to respond to an external command. This leads to uncertainty concerning intent and to unreliability of related performance measures. The difficulty is especially serious in measurements involving young children, and affects speed of response particularly. We advocate observation of the quality rather than the rate of activity as a more reliable measure of performance.

OPERATOR MODELING

Particularly in reference to the design of control systems involving human operators, it has been found necessary to develop mathematical models of the human operator. Such models need not be particularly mysterious and need not involve complex mathematics. Indeed, for this paper no mathematics at all is needed; instead, the idea of a human operator as a building block in analysis of a complex system is employed.

In Figure 1 an elementary block diagram of a human-machine system is given to illustrate this concept. In response to an external input Θ_i, the human operator generates an output x that becomes the input to a mechanical device, causing an output Θ_o. Ideally Θ_o, should equal Θ_{0h}, and the human operator will use feedback concerning the status of Θ_o to adjust x until that condition is met.

The simple system of Figure 1 is appropriate, for example, when an antiaircraft gunner (with now obsolete manual equipment) attempts to correct his aim in order to destroy an attacking aircraft. In that case, because of unquestionable motivation and extensive training, timing of the operator's response would yield valid information. A less militaristic example might involve performance by an elite athlete in priate, however, in that it provided the basis for the original research on human operator performance.

However, this system does not represent adequately the voluntary use of a prosthesis—except perhaps in gunnery. To develop the critical difference, let us continue for the present with the gunnery analogy. If the immediate threat of the attacking aircraft is removed and replaced by a less stressful peacetime exercise in firing at a target drone, a new variable is introduced within the box labeled "human operator." This variable has to do with the delay or possibly with other differences between the external "command" and the operator's "intent." Possibly, out of sheer perversity or in order to irritate a superior, the gunner will deliberately miss the target. To take such a possibility into account, the diagram may be modified as shown in Figure 2 where the "noise" (n) is any input, unrelated to Θ_i or Θ_o, which affects x. In the gunnery example, n might be related to a past grievance or to present distractions.

These comments originate from research supported in part by grants from Health and Welfare Canada, the Natural Sciences and Engineering Research Council, the Canadian Imperial Bank of Commerce, and the War Amputations of Canada.

Address correspondence to Prof. R. N. Scott, Director, Bioengineering Institute, University of New Brunswick, P.O. Box 4400, Fredericton, New Brunswick, Canada E3B 5A3.

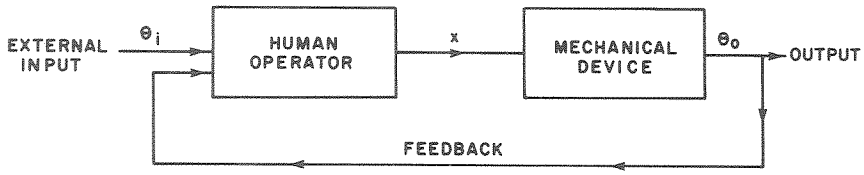

FIGURE 1
Block diagram of elementary man-machine system.

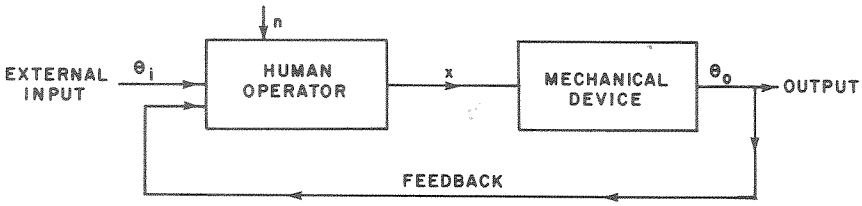

FIGURE 2
Block diagram of man-machine system with noise input.

This is the simplest possible representation of the fact that human behavior is not simply an automatic response to a specific input. There are several stages to this response, all of which become obvious when pointed out. The operator must *1)* recognize the input signal Θ_i, *2)* interpret this signal, *3)* decide on the most appropriate response, and *4)* execute that response. Distractions, inattention, or other factors may affect all four of these stages. The effect is that the experimenter cannot be assured that measurement of x as it relates to Θ_i and Θ_o will represent anything like the operator's best response. To measure that would require a knowledge of internal decisions that are not accessible to the experimenter. Furthermore, attempts to create incentives to fast performance may lead to reduced accuracy. Aside from creating situations of extreme stress (attacking aircraft), this leaves the experimenter stymied.

CLINICAL TESTS

This basic notion is well known, in a different guise, to clinicians. Even the most junior clinician learns quickly that a patient's verbal response to questions may involve a range of "noise" elements, from slight misunderstanding to a deliberate evasion, and that response to requests may be less than whole-hearted cooperation. What then can be done?

First it is essential that the patient (operator) be convinced to cooperate in whatever task is presented. Various strategies, including making the task enjoyable and offering rewards for success, are used. The researcher must realize that experimental data will reflect those inducements as well as the patient's capability.

Second, one should avoid attempting this type of measurement with patients, such as children, who may not be able to comprehend fully the researcher's instructions or who may not perceive the importance of a rapid and accurate response. If comprehension is not reliable, the resulting data are invalid.

AN ALTERNATIVE

An alternative that is more nearly feasible with children and is advantageous generally is to depart totally from the "command-response" format. Rather, one can observe the patient's behavior in a situation where an attractive task is available and may be undertaken at the patient's pleasure. Here "intent" will generally be evident from the patient's initiation of the task, and one may be able to deduce from subsequent events whether the intent continued uninterrupted. The task should be simple or consist of identifiable simple elements to facilitate scoring.

Even in this context, unless the task contains an explicit time element or unless some real urgency is present in the situation, measurement of the time taken to execute the task is at best risky. It requires a further assumption, not only that the patient wishes to complete the task but also that he or she wishes to do this as quickly as possible. Thus assessment of performance should be based on such measures as spontaneity and skill rather than speed.

It is relevant to note that normal human activity is not characterized by attempts to achieve maximum speed, and that the objective of rehabilitation effort is restoration of a degree of normalcy.

CONCLUSION

In determining functional proficiency one is unwise to assume that the patient shares the experimenter's objectives. Observation of spontaneous activities is a more reliable method of estimating actual ability. Speed is not a reliable indicator of proficiency unless it is an inherent element of the task.

R. N. Scott, D.Sc.; E. R. Sandson, O. T.; and P. A. Parker, Ph.D.
Bio-Engineering Institute and
Department of Electrical Engineering,
University of New Brunswick, Fredericton,
New Brunswick, Canada E3B 5A3

The Fallacy of Qualitative Performance Tests

To the Editor:

Thank you for the opportunity to help initiate this important discussion. "The Fallacy of Timed Functional Tests" by Scott et al. addresses a fundamental issue in empirical rehabilitation research: the reliability and validity of performance data. Scott et al. have provided valuable stimulation to rehabilitation professionals: assessment of human performance is complex and fraught with difficulties that have been recalcitrant to efforts by some of the best workers in our field.

From the outset, I must categorically refute the statement that quantitative data, such as timed performance values, are inherently better (or worse) than qualitative data, at least with respect to assessment of disabled individuals and particularly with regard to amputee and prosthesis performance. My position is easy to substantiate: the literature is equally silent on the reliability and validity of timed performance tests as on the usefulness of naturalistic "[observation] of the patient's behavior." Having just completed a study of the reliability of observational analyses, I am less confident than Scott et al. regarding their potential (7).

Scott el al. conclude that "observation of the quality rather than the rate of activity [is] a more reliable measure of performance." *Fallacy*, the term used in the paper's title, indicates invalidity in scientific or epistemological terms. Theoretical arguments—no empirical data—are given to support the contentions. The position I argue is that to say "timed tests are fallacious" is insufficient. It is incumbent upon the agonists of such a viewpoint to *1)* provide evidence of empirical data that supports the contention that timed tests are unreliable and fallacious; and *2)* provide data indicating that qualitative observations are indeed more valid than timed data: viz, that qualitative observations are measured with a minimum of systematic bias and random rater error. Most importantly, I argue that the projected use of a test should be the final arbiter in the choice of assessment techniques.

TEST ASSUMPTIONS

Functional assessments are tools to decrease ignorance, and as such must obey the rules of epistemology. The assessment battery and its constituent tasks must be measurable, scalable, reliable, and valid (6). That a task must be measurable to permit the test to be scored is self-evident. Scalability refers to the scores' arrangement along some dimension to construct a measuring device. Thus, distance is measurable, and one can construct a "yardstick" by arranging 1-inch increments along a standard 3-foot scale. Reliability and validity refer to the consistency and veridicality, respectively, of the test results. Validity assumes reliability; reliability assumes scalability; scalability assumes measurability.

Under optimal conditions, measures are unobtrusive: their scale corresponds isomorphically to gradients that naturally occur in life, and the scores attributed to an event are consistently assigned. Furthermore, under optimal conditions, the scores represent the range of activities that are performed in daily living.

It is important not to confuse concept or item veridicality with scoring reliability. While it is true

that all tests assume willingness to participate by the subject, to assert that measures of spontaneity and skill should be the sole basis of performance assessment is a non sequitur. Distractions, inattention, lack of motivation, and other factors noted by the authors grant no more clemency to qualitative than to quantitative performance ratings. Event timing certainly conforms to the four test requirements noted previously with the possible exception of validity (as Scott et al. quite properly note).

Because Scott et al. offer no examples of the qualitative scales to which they refer, it is not clear that the proposed concepts, spontaneity, and skill satisfy the most basic of test requirements, i.e., measurability and scalability. Although the concept of performance spontaneity is an attractive one, it is difficult to envision a scale that could sufficiently capture its degree of presence or absence along a single dimension. One can only hope that such an urgently needed scale will soon be published, with empirical data on its reliability and validity.

TEST PURPOSES

The choice of test scale and items depends on and ultimately determines the use of the test. If aesthetics is the domain of measurement, qualitative observations may be invaluable, if one can establish that the raters hold views representative of population values. If, however, the test is intended to assess efficiency of prosthetic terminal device use, then timing the amputee on a battery of tasks that compel its use should provide the most realistic estimate of functional proficiency. Functional assessments are usually intended to reflect performance or some aspect of performance in real life.

Because tests of performance quality are essentially projections of the judges' psychological perceptions, investigators should carefully assess the premises of any qualitative scale. Qualitative scales cannot collect zero-order or life-event data. By their nature they intermix data collection and interpretation. If, for example, the purpose of the investigation is to determine the attractiveness or beauty of terminal device use, a panel of Olympic gymnastic-contest judges might be employed to assess the amputee's performance. (Particularly fresh in memory is the influence of each panelist's country of origin on qualitative assessment reliability during recent Olympic judgings.) If the purpose of testing is to determine prosthetic usefulness in efficiently accomplishing self-care or other activities of daily living, a dimension other than solely observer-designated aesthetics should be employed. Qualitative tests should also account for cultural, psychological, and socioeconomic differences between the subject's and the rater's backgrounds.

If the object of rehabilitation efforts is to "restore a degree of normalcy" to disabled persons, criteria such as operation of the prosthetic device at a rate satisfactory to their employers and to themselves should reveal the success of these efforts. Practical examples of rate-dependent activities include performance of vocational activities or attempts to cross a street in time with the traffic light.

EMPIRICAL DATA SUPPORTING TIMED TESTS

Resolution of the issue of timed versus qualitative assessment of performance is not necessarily dichotomous, nor is it problematic only in the field of rehabilitation. Standardized performance tests for nonhandicapped children also suffer from the difficulties of "command-response format" (1, 2, 5), combining timing and qualitative rating scales to assess motor performance of children without disabilities. These tests have been extensively reported, and timing has been shown to be a necessary component. Both quality and timing of performance have been employed in upper-limb terminal device assessments, although the reliability and validity of such scores have not been reported (4).

Keith (6) has extensively reviewed the literature on performance testing of disabled persons and has shown that lack of agreement among rehabilitation professionals for measuring function is an important inhibition to communication in this field. Perhaps the primary inhibition to more widespread use of qualitative functional assessments is that such data are typically regarded as "soft" judgments with little scientific merit (3). This in turn is usually the result of investigators failing to appreciate the need for inclusion of items that can be both performed reliably and scored reliably.

PERFORMANCE RELIABILITY

The issue most forcefully elucidated by Scott et al. is the inabililty of timed tests to assure high motivation. Presumably, however, lack of high motivation would only be a problem when attempting to determine maximal performance between subjects. If two treatments were compared with subjects, the test need only assume equal motivation within the subjects in response to experimental conditions.

Performance inconsistency has classically been addressed by specifying a priori definitions of the

methods for counting and attributing events that occur during the clinical trial (8). Thus, the argument that qualitative assessments of spontaneity and skill are reliable would seem to hinge on each rater's ability to ignore events that occur during periods of the subject's "inattention," scoring only those that fit the rater's notions of the "child's true intent." Such a performance test, if it could be made reliable while using tasks that isomorphically reflect function outside the test environment, would indeed be a significant contribution to the field of rehabilitation.

In the absence of widely accepted, valid assessment batteries, it is incumbent upon each investigator to operationalize and then report test attributes such as specification of when the test period begins and ends, the initial and final positions of the rater, subject and test items, and relevant portions of the performance test. The rater must also have a priori rules to determine when the performance should be judged invalid and when the test should be repeated. Such operational definitions would enhance scoring and performance consistency, while simultaneously facilitating replication of the study. The latter, of course, is a basic tenet of scientific knowledge specifically and of epistemology generally.

Finally, purely statistical considerations bear upon the decision to use timing or qualitative assessments. Event duration by its nature is a "ratio" scale: duration has a naturally occurring zero point and equal intervals between successive scores that correspond isomorphically to reality time-to-completion ratios. Qualitative scales, on the other hand, must carefully eliminate the rater's bias during the inevitable analysis and interpretation that observers will project on the scales during data collection. Furthermore, qualitative ratings must be forced to comply with the real-life event, using a rank-order scale extending from lowest quality to highest quality of performance; in addition, such scales usually lack a naturally occurring zero point. Thus, time is typically analyzed with more powerful and more discriminating parametric statistics, whereas qualitative scales are frequently relegated to less powerful nonparametric or rank-order statistics.

CONCLUSION

The literature is deplorably meager to those seeking standardized, reliable, and valid performance tests, particularly in the area of prosthetic functional assessment. Therefore, before publishing the results of any test, it is incumbent upon investigators to discuss the limits of their data and to provide empirical evidence of the performance and scoring reliability, as well as the relationship of the test results to external criteria that indicate test validity.

Unidimensional measurement of performance, particularly narrowly focused scales such as time-to-completion or attractiveness of a person's function with a prosthesis, can only limit otherwise meaningful research efforts. In emphasizing spontaneous use of orthopedic appliances, Scott et al. have proposed a valuable dimension on which to base judgments of rehabilitation efforts. It should be clear, however, that uncontrived test environments are not antithetical to well-defined standardized or timed tests.

I hope this letter will stimulate others to join in a productive and ongoing discussion of these points in future issues of this journal.

David E. Krebs, M.A., P.T.
Associate Research Scientist,
Prosthetics and Orthotics
New York University Post-Graduate Medical School
317 East 34th St., New York, NY 10016

REFERENCES

1. AYRES AJ: Southern California sensory integration tests. Los Angeles, CA: Western Psychological Services.
2. BRUININKS RH: Bruininks-Oseretsky test of motor proficiency. Circle Pines, MN: American Guidance Service.
3. FEINSTEIN AR: Clinical biostatistics: XLV. Hard science, soft data, and the challenges of choosing clinical variables in research. *Clin Pharmacol Ther* 22: 485–498, 1977.
4. FISHMAN S, KAY H: Acceptability of a functional-cosmetic artificial hand for young children, part II. *Artificial Limbs* 8(2): 15–27, 1964.
5. FRANKENBURG WK, DODDS JB: The Denver developmental screening test. *Pediatrics* 71: 181–191, 1967.
6. KEITH RA: Functional assessment measures in medical rehabilitation: Current status. *Arch Phys Med Rehabil* 65: 74–78, 1984.
7. KREBS DE, EDELSTEIN JE, FISHMAN S: Reliability of kinematic observational gait analysis. *Phys Ther* 65: 1027–1033, 1985.
8. SACKETT DL, GENT M: Controversy in counting and attributing events in clinical trials. *N Engl J Med* 301: 1410–1412, 1979.