ous forms of goniometers are used.[15–18] Boone et al[18] found that the reliability of upper extremity measurement was greater than for the lower extremity. Also they found that one measurement was accurate, whereas Low's study[17] indicated that an average of several measurements should be performed. Recently, Clapper and Wolf[19] evaluated the reliability of the goniometer and an electronic instrument, the Orthoranger. The Orthoranger was in a sense used as the gold standard, having demonstrated only a 2-degree variance over 128 degrees.[20] It is interesting that the standard goniometer demonstrated greater intraclass correlations and confidence levels than the Orthoranger for all movements except lateral hip rotation. Except for hip adduction and knee extension, a positive relationship was shown between instruments.

With regard to different instruments, Hamilton and Lacherbruch[21] found equal accuracy and reliability when comparing three different goniometers (dorsal, universal, and pendulum). They measured finger joint angles and standardized the procedure. Rothstein et al[22] compared three different goniometers (large metal, large plastic, and small plastic) for reliability in measuring passive elbow and knee movement. They found high interdevice reliability with an intraclass correlation coefficient of >0:91.

Some common concerns with regard to variability in measurement are the starting position, speed and smoothness of motion, and the size and contour of the extremity being measured. Another important factor is that if the testing position is not standard, the measurement will differ considerably. For example, knee flexion in the prone position will always be less due to the tension in the rectus femoris; whereas, measured supine, the hip may be flexed to eliminate this factor (if appropriate). Finally, it is known that there is variability among joints. For example, knee extension demonstrates a greater margin of error than many other joint measurements. Hellebrandt et al[16] found wrist flexion, medial rotation of the shoulder, and shoulder abduction difficult to position and measure.

Most studies have been performed on active ROM. Passive ROM is considered less reliable because of the need to control the body segment and measure at the same time. Wagner[23] found that variability of passive movement was generally higher than the variability with active movement when evaluating forearm pronation and supination. Bird and Stowe[24] found that error in measurement for passive wrist movements was greater than for active wrist movements.

MUSCLE TESTING (STRENGTH)

Muscle strength testing is a classic example of how a procedure that has been demonstrated to have some intratester or intertester reliability unfortunately carries little validity with most patients. Several studies indicate that intraexaminer and interexaminer agreement on test and retest occurs 40% to 75% of the time.[25–27] When agreement within one full grade is measured, intraexaminer reliability was 95% with interexaminer reliability at 90%. As will be demonstrated, this is misleading due to the extremely poor ability to correlate "true" muscle strength with suspected

levels based on manual testing. Although this is known by most physicians, it still remains a common procedure used in the musculoskeletal evaluation of patients. The question is, perhaps, what is the purpose of the test, or rather what information is being sought? And more important, how valid is the use of the information that is gained?

Classically, muscle testing is used to determine weakness. This weakness is then differentiated based on the suspicion of whether it is neurologic or myopathic. The most accepted grading system for test results was, in fact, generated from a pathologic group, i.e., polio patients.[28] Although fairly reliable regarding reproducibility and accuracy for patients with debilitating weakness, testing of other groups such as a control group of normal patients and a group of patients with other causes (such as low back patients with nerve root involvement) has not been confirmed as equally reliable.

The original grading system uses five grades. Another grading system has been suggested that would divide responses into good, fair, and poor categories associated with numerical equivalents of 75%, 50%, and 25% of normal strength, respectively.[25, 29, 30] Although the fair grade is the most accurate, when actually compared with mechanized isokinetic testing, a range of 6% to 32% was found rather than the suggested 50%.[29] In addition, patients were often rated as normal by manual muscle testing when compared with other controlled, quantified tests.[31, 32] These tests indicated that the patients had a 50% loss of strength. A good rating has been given by examiners on muscles with strength as low as 8% of the expected normal.[33]

Even more revealing are the findings of studies evaluating agreement between extremities. Examiners rated the stronger limb as weak 20% of the time.[34] This occurred most often when the quantitative strength difference was less than 10%. When a comparison between extremities resulted in a good rating for one limb and normal for the other, the actual difference in strength was 35% or more.[30] Discriminating differences in strength of less than 25% is difficult even for trained examiners.

It is important to remember that these tests are for generic groups of agonists such as adductors or internal rotators. Assessing the function of an individual muscle adds a complex variable to the already questionable validity of muscle testing. Although electromyographic (EMG) studies have attempted to determine the position in which an individual muscle is most active, total isolation of that muscle is impossible. In other words, all of the variables that decrease the validity of group muscle testing make individual muscle testing (for the purposes of determining the degree of weakness) even less valid.

Some of the variables to consider include the following:

- Variation in the technique of application. Slight changes in body position may affect the patient's ability to contract fully. This is not only true of the joint being tested but for proximal and distal joints as well. An example of this phenomenon is the variation in strength of the internal tibial rotators with change in the position of the hip. This occurs even though the knee is held constant at a 90-degree angle.[35]

- The velocity of the test maneuver.[36, 37]
- Gravity.[38, 39]
- Patient instruction and familiarity with the test.[40]
- Time of day.[41]
- Painful inhibition.
- Patient willingness.
- Subject or ambient noise.[42]
- Inability to properly stabilize the muscle origin. Inadequate stabilization may prevent the maximum contraction of muscles used for a distal movement.[43, 44] Inadequate stabilization may also allow stronger proximal muscles to be used, leading to what is often termed recruitment.[45, 46]

Obviously standardization of testing would help with inter/intraobserver reproducibility. The examiner should be aware of the findings of Nicholas et al.,[47] which suggest that testers mentally integrate the amount of force and the time of application in arriving at a rating. It appears that if an equal or greater force is generated by one limb compared with another, it may be judged as weaker if it is shorter in duration. Also findings by Hsieh and Philips,[48] using a computerized dynamometer, suggest that a "patient-initiated" method of testing produced greater intratester reliability. This method has the patient initiate the contraction with the instruction "push against me as hard as you can." The tester applies additional force after perceiving a maximal attempt by the patient until a "break" is noted.

A final concern is the inferential capability of the clinician armed with muscle testing results. As will be illustrated later in this chapter, this objective data may have no correlation to functional ability, at least in terms of activities of daily living as reported by the patient. Also no one testing approach has been demonstrated to have greater predictive value in this respect when compared with other approaches.[49] On the other hand, more sophisticated testing such as isokinetic measurement may help in predicting those more likely to be injured in sport and occupational settings.

Perhaps muscle testing will be found to have more value in reproducing a patient's pain complaint where a quantitative discrimination is not needed. Also the operational definition of muscle testing will include stretch testing. Therefore, when a muscle is suspected as the tissue involved, both contraction and stretch will be used to determine not only strength deficits but also reproduction of pain. Stretch may then be used less as an assessment of quantitative measurement with inferences as to function and more as a further attempt to reproduce a presenting complaint.

PALPATION

Palpation is often used as an assessment of tenderness, hypertonicity, hypomobility, hypermobility, and measurement (iliac crest height, degree of movement, etc.). Tenderness location has not been found to be very reliable among examiners. Bony tenderness is more easily agreed upon