The Reliability of Upper- and Lower-Extremity Strength Testing in a Community Survey of Older Adults

Kenneth J. Ottenbacher, PhD, OTR, Laurence G. Branch, PhD, Laura Ray, MPA, Vera A. Gonzales, PhD, M. Kristen Peek, PhD, Martha R. Hinman, EdD, PT


Objective: To examine the stability (test-retest reliability) of strength measures in older adults obtained by nontherapist lay examiners by using a hand-held portable muscle testing device (Nicholas Manual Muscle Tester).

Design: A prospective relational design was used to collect test-retest data for 1 male subject by using 27 lay raters who completed intensive training in manual muscle.

Setting: Data were collected from older Mexican-American adults living in the community.

Participants: Twenty-seven lay raters who completed intensive training in manual muscle testing for a field-based assessment and interview of older adults and 63 Mexican-American subjects completing wave 4 of the Hispanic Established Populations for the Epidemiologic Study of the Elderly.

Interventions: Training involved reviewing a manual describing each testing position followed by approximately 6 hours of instruction and practice supervised by an experienced physical therapist. Lay raters then collected test-retest information on older Mexican-American subjects.

Main Outcome Measure: Stability (test-retest) for a portable manual muscle testing device.

Results: Intraclass correlation coefficients (ICCs) were computed for the 27 lay raters examining 1 male subject (2 trials) and 12 lay raters assessing 63 older Mexican-American adults (1 practice and 2 trials recorded). The ICC values for the first 27 lay raters ranged from .74 to .96. The ICC values for the latter 12 lay raters ranged from .87 to .98. No differences were found in ICC values between male or female subjects.

Conclusions: Stable and consistent information for upper- and lower-extremity strength was collected from the older adults participating in this study. The results suggest reliable information can be obtained by lay raters using a portable manual muscle testing device if the examiners receive intensive training.

Key Words: Aging; Hispanic Americans; Muscles; Rehabilitation; Reliability and validity.

© 2002 by the American Congress of Rehabilitation Medicine and the American Academy of Physical Medicine and Rehabilitation

The components of the disablement process (pathology, impairment, functional limitation, disability) have been widely described in the rehabilitation and aging literature. Previous studies of the disablement process have focused on the identification of risk factors, the clarification of their influence on disability, and the relationship among impairments, functional limitations, and disability. The disablement process in older at-risk or disabled adults is frequently examined by using physical performance tests and strength measurements. For example, Jette et al examined the interrelationships among measures of impairment, functional limitation, and disability in older community-dwelling adults. They used field-based measures of upper- and lower-extremity strength collected with a hand-held muscle testing device (Nicholas Manual Muscle Tester) to help operationalize components of the disablement process.

We hypothesized, based on previous research on the disablement process, that muscle strength influences subsequent disability in older adults. To address this hypothesis, we have undertaken a study to examine the influence of muscle strength on the process of becoming disabled among older Mexican-American adults. The first step in this research program is to establish the reliability of performance testing and strength measurement used to evaluate components of the disablement process.

Strength measures in previous community-based investigations of older adults have usually been collected by trained rehabilitation professionals such as physical or occupational therapists. Advances in rehabilitation engineering and technology have produced portable instruments that are designed to provide precise and objective assessment of muscle strength. These devices are designed to provide objective estimates of muscle strength that do not rely on the skill and judgment of a rehabilitation therapist or trained professional using traditional manual muscle testing procedures. Little information is currently available regarding the psychometric properties of the new generation of hand-held manual muscle testers, particularly when used by nontherapist lay examiners.

The goals of the present study were (1) to examine the stability (test-retest reliability) of the Nicholas Manual Muscle Tester in evaluating upper- and lower-extremity muscle strength in older adults and (2) to determine if consistent recordings could be obtained using trained raters who were not rehabilitation (physical or occupational) therapists.

METHODS

Instrument

We used the Nicholas Manual Muscle Tester to assess muscle strength in the fourth wave of the in-person data collection for the Hispanic Established Populations for the Epidemiologic
Study of the Elderly (EPESE). The Nicholas Manual Muscle Tester is a hand-held device for objectively quantifying isometric muscle strength. The peak strength (in kilograms) required to break an isometric contraction is measured as the examiner applies force against the subject. This muscle tester is designed to be used with larger muscle groups of the upper and lower extremity. A load cell in the device provides digital output ranging from 0.0 to 199.9kg (equivalent to approximately 440lb). The unit is placed between the examiner’s hand and the limb being tested (see detailed description below).

**Procedures**

The study involved 2 phases. In the first phase, 27 lay raters serving as interviewers in the fourth wave of data collection for the Hispanic EPESE assessed the strength of 1 male adult subject. There were 13 male and 14 female interviewers; the average age for the men was 50 years and for the women, 42 years. All 27 interviewers were hired to collect data for the Hispanic EPESE investigation and none were licensed or trained rehabilitation professionals. Detailed information regarding the complete Hispanic EPESE data collection protocol is available in previous publications. During the training, each rater completed 3 measurements (1 practice followed by 2 trials). Only the 2 trials were recorded. Testing was performed as part of a 2-day structured training of interviewers in preparation for the fourth wave of the Hispanic EPESE. On the first day of training, a physical therapist with 20 years of experience went through the training manual (available on request), referred to the pictures provided in the training manual, and showed each test on the male subject. After the demonstration, the raters broke into small groups (4–5 persons) to practice on each other. The interviewers referred to the manual, and members of the research team were available to answer questions and observe the practice testing. Members of the research team and the physical therapist provided feedback or adjusted testing positions when necessary. This training lasted approximately 3 hours.

On the second day of training, each of the raters tested the same single male subject who performed the following movements: hip abduction, hip flexion, knee extension, shoulder abduction with arm outstretched and parallel to the floor, and shoulder abduction with arm at his side. We adapted a protocol described by Jette et al in testing older adult subjects in the community. The male subject’s right (dominant) arm and leg were tested. The raters tested the subject over a 3-hour period that was part of the training program. The subject was tested by 2 rater pairs and then had a brief rest period (2–3min) before testing by the next pair of raters. Longer breaks of 15 to 30 minutes were taken twice during the training-testing period.

The male subject was seated in a straight-backed chair with his feet flat on the floor. For each assessment, the raters positioned themselves so their forearms and the manual muscle testing device were perpendicular to the muscle group being tested. A member of the research team silently read each score and provided it to the recorder so that no interviewer or rater would be aware of previous scores for the male subject. The adult male subject, tested by all 27 raters, was selected based on prior tests that indicated his strength levels exceeded those expected from the Hispanic respondents who would be aged 70 and older during this wave of interviewing.

The second phase of the study involved data collected from the first 63 subjects interviewed and tested as part of the larger Hispanic EPESE investigation. Data on these subjects were collected by 12 of the interviewers who participated in the training session described previously. The 63 subjects met the inclusion criteria for the Hispanic EPESE study and were also participants in the fourth wave of data collection. All subjects were living in the community. Forty-four percent rated their overall health as either good or excellent. Approximately 28% reported a history of diabetes, 4% reported a previous stroke, 6% a heart attack, and 7% a previous diagnosis of cancer. Additional descriptive information is included in the next section. The same administration protocol followed in collecting test-retest information for the training session was used to gather strength measurements from these 63 subjects. Each of the 63 subjects was offered a practice opportunity (which was recorded) followed by 2 assessment trials. The right arm and leg were tested for all subjects unless medically prohibited (eg, pain, previous surgery). Responses from all 3 trials were recorded.

The research procedures in the Hispanic EPESE study were reviewed and approved by the appropriate university institutional review board.

**Data Analysis**

Data were analyzed by using the intraclass correlation coefficient (ICC) approach based on generalizability theory. The ICC expresses measurement error and agreement as the relationship between true and observed variance. The coefficients are derived from an analysis of variance model. The ICC is widely used as a method to examine inter- and intrarater agreement in the biomedical and behavioral sciences. Model (3,1), referred to by Shrout and Fleiss as the fixed effects model for the ICC, was used to compute the ICC value for stability (intra-rater reliability). This model assumes time (first vs second vs third administration) as a fixed effect. ICC values were calculated separately for the training component of the study that included 27 different raters examining the same subject over 2 trials and for the Hispanic EPESE component of the study that involved 12 raters examining 63 older Mexican-American subjects with 1 practice trial and 2 assessment trials.

### RESULTS

**Phase 1**

The ICCs for the 5 strength dimensions with the single male subject and 27 examiners are presented in table 1. Three ICC values were above .90; one was between .80 and .90, and one was at .79. For the female examiners, 3 ICC values were above .90 and the other 2 were between .80 and .90. For the male examiners, 1 ICC value was above .90, 3 were between .80 and .90, and the lowest was .74. Overall, the ICCs were high, suggesting good stability of ratings across the 27 lay interviewers trained to collect data for the Hispanic EPESE investigation.

**Phase 2**

As noted earlier, phase 2 of the study involved examining the stability of the assessment across 1 practice trial and 2

<table>
<thead>
<tr>
<th>Table 1: ICC Values for 27 Raters Testing 1 Subject Across 5 Positions for Trials 1 and 2</th>
<th>All Raters (N=27)</th>
<th>Men (n=13)</th>
<th>Women (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hip abduction</td>
<td>.96</td>
<td>.88</td>
<td>.95</td>
</tr>
<tr>
<td>Hip flexion</td>
<td>.94</td>
<td>.94</td>
<td>.94</td>
</tr>
<tr>
<td>Knee extension</td>
<td>.84</td>
<td>.84</td>
<td>.83</td>
</tr>
<tr>
<td>Shoulder abduction, arm at side</td>
<td>.90</td>
<td>.81</td>
<td>.93</td>
</tr>
<tr>
<td>Shoulder abduction, arm outstretched</td>
<td>.79</td>
<td>.74</td>
<td>.85</td>
</tr>
</tbody>
</table>

NOTE. All testers provided 2 assessments of each position; all assessments were done on the subject’s right (dominant) side.
measurements for 63 older Mexican-American subjects participating in the Hispanic EPESE investigation. These 63 subjects were assessed by 12 of the original 27 raters. The range of subjects tested by the 12 examiners was 1 to 15; the mode was 3 subjects. Basic descriptive information for the 63 subjects from the Hispanic EPESE investigation appears in Table 2. The ICC values for the 63 subjects are presented in Table 3. The table contains ICC information for 3 assessments: the practice and trials 2 and 3. The ICC values for trials 2 and 3 ranged from .87 to .98. Figure 1 includes the scatterplots showing the relationship between trials 2 and 3 for the 5 testing positions on the 63 Hispanic EPESE subjects.

By using the first trial for the 63 subjects, we examined muscle strength by gender. As expected, statistically significant differences in muscle strength were noted between men and women for the 5 positions tested: hip abduction (t = 5.43, P < .000), hip flexion (t = 3.67, P < .001); knee extension (t = 3.27, P < .002); shoulder abduction, arm at side (t = 5.60, P < .000); and shoulder abduction, arm outstretched (t = 3.87, P < .000). In all of these comparisons, the male subjects recorded higher muscle strength values than the women. There was no statistically significant difference in age between male and females subjects (t = .42, P = .67). There were low to moderate correlations between age and muscle strength across the 5 positions. The correlations ranged from r equal to -.12 for knee extension to r equal to -.32 for shoulder abduction, arm at side.

**DISCUSSION**

The goals of this study were to examine the stability (test-retest reliability) of the Nicholas Manual Muscle Tester in evaluating upper- and lower-extremity muscle strength in older adults. We were also interested in determining if consistent recordings could be obtained by using trained examiners who were not rehabilitation (physical or occupational) therapists. The results suggest that the Nicholas Manual Muscle Tester can produce consistent measures of upper- and lower-extremity strength in older community-dwelling adults. The majority of the ICC values for trials 2 and 3 in the 63 older Mexican-American subjects were greater than .90, indicating excellent measurement consistency over trials. In discussing interpretation of the ICC, Portney and Watkins17 state that “As a general guideline, we suggest that values above .75 are indicative of good reliability, and those below .75 poor to moderate reliability.”

The results of our investigation suggest that lay raters who are provided systematic training, including written documentation with illustrations, can reliably collect strength information by using a portable manual muscle testing device. The ICC values for the 27 lay interviewers ranged from .74 to .96. In testing the 63 Hispanic EPESE subjects, the lowest ICC value for the lay raters was .74 for shoulder abduction with the arm outstretched. This value was reported for the male raters.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total (N=63)</th>
<th>Men (n=23)</th>
<th>Women (n=40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (y)</td>
<td>70.51±4.73</td>
<td>70.17±4.31</td>
<td>70.70±4.99</td>
</tr>
<tr>
<td>Mean education (y)</td>
<td>4.19±3.88</td>
<td>4.45±3.91</td>
<td>4.05±3.90</td>
</tr>
<tr>
<td>Married (%)</td>
<td>57</td>
<td>78</td>
<td>45</td>
</tr>
<tr>
<td>Native born (%)</td>
<td>60</td>
<td>48</td>
<td>67</td>
</tr>
<tr>
<td>English interview (%)</td>
<td>22</td>
<td>30</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 2: Descriptive and Demographic Information for 63 Subjects in the Hispanic EPESE Study

Abbreviation: SD, standard deviation.

Previous research18-20 on the reliability of strength testing has produced mixed results depending on the sample tested, the device used for testing, and/or the muscle group (position) assessed. The existing research suggests that investigators cannot assume that strength testing in older adult subjects will produce consistent results.21 We showed a high rate of agreement across trials for trained lay raters testing a standard subject and also for multiple lay raters testing 63 older adult subjects. The test-retest reliability for the 5 positions we examined was excellent.

The results indicate that upper- and lower-extremity strength measures can be reliably obtained in older adults by trained raters by using the Nicholas Manual Muscle Tester. The strength-related information will be useful in exploring the relationship among the components of the disablment process in older Mexican-American subjects participating in the Hispanic EPESE.

The stability of the measures obtained in our study must be qualified based on the following limitations. The lay examiners involved in this study received short-term extensive training in administering the strength testing and other performance-based assessments included in the Hispanic EPESE investigation. The examiners participated in a 2-day training seminar involving several different professionals. All the lay examiners were instructed in the strength testing methodology by an experienced physical therapist. The examiners were provided multiple opportunities to practice on each other and on the physical therapist who provided training. In addition, the lay examiners were given a detailed training manual including written directions and pictures of each of the testing positions.

The sample size for the 2 phases of the study was not large (27 for the lay raters, 63 for the Hispanic EPESE subjects). The reliability of raters may vary depending on the characteristics of the sample, and our results do not guarantee the same degree of consistency with a different or more disabled sample. Rater or reliability drift may also occur as the raters test more subjects, and we will continue to monitor the consistency of measurements as additional subjects are tested.

**CONCLUSION**

In discussing the importance of reliability in clinical research, Portney and Watkins17 state, “Reliability is fundamental to all aspects of clinical research, because without it we cannot have confidence in the data we collect, nor can we draw rational conclusions from those data.” The information provided in this investigation suggests that the Nicholas Manual Muscle Tester can repeatedly collect strength information by using trained raters who have received systematic training. The method used in the Nicholas Manual Muscle Tester appears to produce consistent results when tested by trained raters. The reliability of strength testing has been produced for upper- and lower-extremity muscle strength in older community-dwelling adults. The ICC values for trials 2 and 3 in the 63 older Mexican-American subjects were greater than .90, indicating excellent measurement consistency over trials.

**Table 3: ICCs for 63 Hispanic EPESE Subjects, Computed for 5 Strength Testing Positions**

<table>
<thead>
<tr>
<th></th>
<th>N=63</th>
<th>Men (n=23)</th>
<th>Women (n=40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC for trials 2 and 3</td>
<td>.93</td>
<td>.93</td>
<td>.87</td>
</tr>
<tr>
<td>Hip abduction</td>
<td>.97</td>
<td>.98</td>
<td>.95</td>
</tr>
<tr>
<td>Hip flexion</td>
<td>.98</td>
<td>.98</td>
<td>.98</td>
</tr>
<tr>
<td>Knee extension</td>
<td>.97</td>
<td>.94</td>
<td>.98</td>
</tr>
<tr>
<td>Shoulder abduction, arm at side</td>
<td>.92</td>
<td>.96</td>
<td>.87</td>
</tr>
<tr>
<td>Shoulder abduction, arm outstretched</td>
<td>.93</td>
<td>.96</td>
<td>.89</td>
</tr>
</tbody>
</table>

The information presented in this investigation suggests that the Nicholas Manual Muscle Tester can reliably collect strength information by using trained raters who have received systematic training. The method used in the Nicholas Manual Muscle Tester appears to produce consistent results when tested by trained raters. The reliability of strength testing has been produced for upper- and lower-extremity muscle strength in older community-dwelling adults. The ICC values for trials 2 and 3 in the 63 older Mexican-American subjects were greater than .90, indicating excellent measurement consistency over trials.
Fig 1. Scatterplots for 5 positions using strength measurements 2 and 3 (in kilograms) for 63 Hispanic EPESE subjects. (A) Hip abduction; (B) hip flexion; (C) knee extension; (D) shoulder abduction (arm at side); (E) shoulder abduction (arm outstretched).
Muscle Tester can provide consistent information on upper- and lower-extremity strength in older adults when used by trained lay examiners.

References


Supplier

a. Model 01160; Lafayette Instrument, PO Box 5729, 3700 Sagamore Pkwy N, Lafayette, IN 47903.