Interrater Reliability of Students Using Hand and Pinch Dynamometers

Debra Lindstrom-Hazel, Andrew Kratt, Laura Bix

KEY WORDS
• hand strength
• muscle strength dynamometer
• pinch strength
• reproducibility of results
• work capacity evaluation

OBJECTIVE. We sought to analyze the interrater reliability of occupational therapy students using standard hand-strength instruments.

METHOD. Seventy-three healthy students, faculty, and staff members from a midwestern university participated in this study. Each participant completed three trials for both grip and pinch strength bilaterally. Two student raters independently scored each trial.

RESULTS. The interrater reliability of the grip and pinch strength trials was determined using an intraclass correlation coefficient (ICC) one-way random effects model (1,1). The ICC for the Jamar dynamometer ranged from .996 to .998 ($p < .05$). The ICC of the pinch gauge ranged from .949 to .990 ($p < .05$).

CONCLUSION. This study demonstrated that occupational therapy students can be reliable raters after being trained and tested for competency. To improve psychometric properties for instruments commonly used in the profession, students can participate in collecting normative data for hand-held dynamometry.


Holm’s 2000 Eleanor Clark Slagle Lecture suggested that all occupational therapists take on the habit of searching for evidence for practice, appraising its value, and presenting it to their clients in an understandable way (p. 575). One way to develop this habit is to require students to cite the evidence on which they are basing their clinical decisions and appraise its value for worthiness of a clinical decision while they are in didactic classes. With these expectations as a backdrop, students at a Midwestern university questioned the validity of the normative data that were used during interpretation of grip and pinch strength for a work capacity evaluation. The most commonly cited norms they could find in occupational therapy text books and from the instrument vendors were more than 20 years old (Mathiowetz, 1985). One hand rehabilitation source (Fess, 2002) recommended that therapists compare one hand to the contralateral extremity rather than using norms, but this recommendation would not be relevant to therapists conducting work or functional capacity evaluations because the intention in that evaluation is to compare the client’s strength to a normative population to determine whether hand strength would be an asset or a limiting factor for that particular person.

The students’ concern about using the 1985 norms included the difference between people’s daily occupations (vocational, avocational, and activities of daily living) in the 1980s compared with people’s daily occupations today. Daily use of computers at home and at work, video games, and small hand-held electronic devices have changed the activity demands and the environmental context for hand use in the past 20 years (Wilson, 1998). Updated norms are imperative to
determine whether the expected hand strength for the U.S. population has stayed the same, increased, or decreased in the past 20 years.

Neuropsychological literature recommends that norms for tests have a lifespan of 15 to 20 years (Strauss, Sherman, & Spreen, 2006), assuming that all other factors remain constant. In psychological testing, selecting an appropriate normative data set is considered a prerequisite for "effective and competent" practice. Choosing a test with an inadequate normative sample is considered as disastrous as choosing a test with poor reliability and validity" (Strauss et al., 2006, p. 44). On the basis of the Strauss et al. (2006) recommendation that norms be revised every 15 to 20 years and the argument that there has been a significant change in activity demands and environmental context over the past 20 years (Wilson, 1998), it is time for an updated grip and pinch norm study.

To have a good comparison for any particular client, norms must be available that match the client’s characteristics as closely as possible on variables such as age, gender, and geographic location and match the administration and scoring procedures of the test used (Strauss et al., 2006). Preliminary work is therefore needed to prepare for a large normative study that could indeed be a valid normative comparison that clinicians throughout the United States could use with confidence. The first step toward conducting this study is to look at the reliability of the raters who would collect such a large data sample. Grip and pinch norms could potentially be gathered by occupational therapy students to get current data from around the country that would be representative and relevant for people in the United States. Before advocating for such a study, however, it is critical to determine whether entry-level occupational therapy students can be reliable raters for grip and pinch strength measurements. In this study, we sought to determine whether occupational therapy students from one midwestern university were reliable raters for grip and pinch strength after a training session and competency testing in the specific administration and scoring procedures.

It is important to establish interrater consistency when administering assessments with more than one data collector. A common design for interrater reliability studies in psychology includes an analysis of a large number of comparison ratings between judges when a battery of tests is administered to a small number of clients by five independent pairs of evaluators (Stringer & Nadolne, 2000). Interrater comparison analysis is needed because interrater differences in scores can produce more than acceptable measurement errors (Edwards, Feightner, & Goldsmith, 1995). Most studies on interrater reliability in rehabilitation rely on professional therapists for data collection (Chen, 2007) or do not specify who reads the gauges that the participants squeezed (Agnew & Maas, 1991; Shechtman, 2004; Sisto, 2007). Edwards et al. (1995) determined that with careful attention to training and the use of standardized administration guidelines, research assistants with no previous background in health care can administer assessments with excellent reliability (intraclass correlation coefficient [ICC] = 0.97–1.00). The research assistants in Edward et al.’s (1995) study were trained in four assessments by an experienced occupational therapist. The study recommended a focus on theoretical and practical training with attention to the standardized assessment guidelines for each assessment. Training specifications and published interrater reliability are important elements for evidence-based decision making.

Mathiowetz (1984) reported the interrater reliability in his norm study to have an ICC of 0.99 for reading a dynamometer and of 0.98 for reading a pinch gauge. Kolber and Cleland (2005) found the ICC of raters testing participants without known impairments to be 0.52 to 0.93 when using hand-held dynamometry. Peolsson, Hedlund, and Oberg (2001) found the interrater reliability ICC to be 0.98 for handgrip using the Jamar dynamometer (Sammons Preston Patterson Medical Division, Bolingbrook, IL) for hand strength. In a review of 18 studies, Bohannon (1999) found the majority of reliability coefficients for hand-held dynamometry to be above 0.7. MacDermid, Kramer, Woodbury, McFarlane, and Roth (1994) found the ICC of interrater reliabilities of grip, lateral pinch, and tripod pinch measurements in patients with cumulative trauma disorders to be above 0.87 using hand therapists as raters. Although Edwards et al. (1995) found non–health care personnel were valid raters, no studies have examined the reliability of entry-level occupational therapy students as grip strength data collectors.

Method

Student Raters

A convenience sample of four occupational therapy students volunteered to participate in this study. Two of the raters were senior-level undergraduate students in their last semester before completing their bachelor’s degree, and two were graduate students who had completed at least one affiliation. This participation fulfilled part of the research requirements for the student raters’ master’s degree in occupational therapy. The four students were assigned to six different rating teams; each rater served on three rating teams (each rater was paired with each of the other three raters) for data collection (see Table 1). Each rating team rated at least 180 data points. This included grip and pinch testing for at least 10 participants with 18 readings per participant—three different tasks three times with two hands.
The participants included a convenience sample of 73 healthy students, faculty, and staff members from a midwestern university. The focus in this study was on the occupational therapy student raters rather than on the volunteer participants because the purpose of the study was to determine how closely the student raters’ readings were to each other. The participants’ scores were only used to compare readings between raters.

Instruments

A Jamar hydraulic dynamometer and a B & L Engineering pinch gauge were used in this study. Both instruments had been calibrated at a registered laboratory less than 6 months before the study.

Procedures

This study was approved by the university institutional review board to ensure ethical compliance with national standards for research.

Student Rater Training

To monitor student raters’ validity, all of the raters were trained and passed a competency test for administering and reading grip strength testing using the Jamar dynamometer and the B & L Engineering pinch gauge following the American Society of Hand Therapists guidelines (Fess & Moran, 1981). These guidelines call for the person to sit in a straight-backed chair with both feet flat on the floor and the shoulder adducted and neutrally rotated. The elbow is to be flexed at 90°, the forearm should be in neutral, and the wrist should be between 0° and 30° extension and between 0° and 15° ulnar deviation. The arm should be held in space rather than supported on an armrest or by the examiner. The dynamometer is given to the person 3 times in each hand, beginning with the left hand each time. The protocol for the B & L Engineering pinch gauge included the same sitting and arm position with the exception that the student rater administering the testing supported the gauge while the participant squeezed to complete the three trials in each position with each hand.

The training and competency test was conducted by Debra Lindstrom-Hazel before data collection began (see the Appendix). We developed special guidelines for reading the instruments to help the novice-level research students standardize reading the gauge. See Figures 1 and 2 for examples of these guidelines, which included always reading the gauge up to the next line if the indicator was between two lines. Before each data collection session, the student raters completed a warmup of at least 10 readings to refamiliarize themselves with the gauges, including where the pounds or kilograms were read and what the unit markings were on each gauge. During this warmup, the student raters compared their readings and resolved any errors to help ensure reliable ratings during the data collection.

One student investigator (Andrew Kratt) observed the two raters who administered the protocol and collected data. He monitored to be sure that the protocol was administered properly and that the raters did not compare their scores during the data collection phase. The data collectors alternated who administered the testing protocol and followed the American Society of Hand Therapists standards (described earlier). Each of the two investigators recorded a score for each trial completed. Both data collectors recorded every trial for each participant tested. Each participant completed a total of 18 trials.

Data Analysis

We used a type (1,1) ICC one-way random effects model in which people effects are random with a 95% confidence interval to obtain a comparison of measurements between

<table>
<thead>
<tr>
<th>Table 1. Student Rating Team Combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Rater</td>
</tr>
<tr>
<td>Team 1</td>
</tr>
<tr>
<td>Team 2</td>
</tr>
<tr>
<td>Team 3</td>
</tr>
<tr>
<td>Team 4</td>
</tr>
<tr>
<td>Team 5</td>
</tr>
<tr>
<td>Team 6</td>
</tr>
</tbody>
</table>

![Figure 1. The dial on the Jamar dynamometer is above the 60 lb mark, so the reading in this case would be 65 lb.](image_url)
An ICC is typically used when comparing the results of two or more raters because it includes a calculation of the reliability index of the measurement error between judges (raters). The ICC is typically the ratio of the variance of interest over the sum of the variance of interest plus error. The ICC provides a more reliable measure of consistency between raters than does a linear relationship correlation statistic such as a Pearson correlation (Shrout & Fleiss, 1979). We calculated separate ICC analyses for the readings on the Jamar dynamometer and the readings on the B & L Engineering pinch gauge because the dials were very different. Each of the six rating teams had a minimum of 180 ratings.

Results

See Table 2 for a compilation of the team ICC results. For the Jamar, the ICC ranged from .996 to .998 between the six teams; for the combined pinch gauge, the ICC ranged from .949 to .990 between the six teams. The difference between the readings of lateral and three-jaw chuck pinches was minimal, so we report the rating comparisons together because the same gauge was read for both positions.

Discussion

The ICC scores are consistent with the findings from Mathiowetz’s (1984) reliability study using similar gauges and with the Edwards et al. (1995) study examining inter-rater reliability for grip strength testing between trained non–health care providers. Although the interrater reliability was slightly higher with the Jamar than with the pinch gauge, the pinch gauge interrater reliability was still extremely high (above 0.90).

The combination of graduate and undergraduate students on rating teams did not seem to make a significant difference in the ratings. These preliminary results suggest that both upper-level undergraduate occupational therapy students and graduate students who have been properly trained and have passed a competency test for this testing have the potential to be reliable raters when using the Jamar hand dynamometer and the B & L Engineering pinch gauge to collect grip and pinch norms.

Limitations

This study had several limitations. First, there was no screen between raters to impede the view of the individual raters’ recordings, but one observer consistently monitored the rating to be sure that there was no discrete consultation between raters during the data collection. The second limitation was the slightly inflated readings when the dial was between the lines on the gauge. This reading procedure (to go up to the next mark when the dial was between lines on the gauge), however, makes the ratings more reliable. In the very early work on this study, we commonly noted rating differences when the dial was between two lines and raters attempted to estimate to the closest marking. Following this procedure could be a problem if a therapist is testing for incremental strength improvements, and it is only recommended when comparisons are needed between multiple raters (i.e., between a work capacity evaluation and norms) to decrease the opportunity for human interpretation and error. The last limitation is that the student raters were not randomly chosen but were from three different tracks of occupational therapy students at one university. We do not claim that all occupational therapy students will be reliable raters, but the results of this study suggest that it is possible for occupational therapy students to be trained to be reliable raters.

Implications

Just as athletes prepare their bodies for an activity (running or swimming), students review their notes before a test, and clinicians review protocols before treating, we recommend that raters review the instruments and gauges before data collection. If clinicians are administering evaluations using different instruments, they might consider taking 20 s to...
familiarize themselves with each instrument before trying to read the dial. The need to familiarize oneself with multiple instruments becomes more of an issue when they have different types of gauges (i.e., Jamar dynamometer and a pinch gauge) with different dials that have both pound and kilogram markings, increasing the reading difficulty for the correct unit of measurement. We recommend warming up with a minimum of five readings on each gauge before initiating readings for norm collection.

This study has implications for future norm collection research. With proper training and competency testing, occupational therapy students have the potential to be reliable raters to collect norms for grip and pinch strength. Involving occupational therapy students in the process of updating norms for commonly used assessments meets the needs of students learning research skills and the need for the profession to have current norms that are valid and reliable for evidence-based decision making. ▲

Acknowledgments
We acknowledge Samantha Harrow, Erin Courttnay, Theresa Lynch, Kelly Sullivan, Nancy Kiddy, Eric Kou, Dr. Carolyn Glogoski, and Michael Chan for their contributions to the early development of the standardized protocol. We also thank Kyle Shearer, Lindsey Carlson, and Emily Myers for data collection assistance. Finally, we thank Cindee Quake-Rapp, Rachel Iott, Stacy Froseth, Tina Caruso, Angela Moran, Javier de la Fuente, Josh Vincent, and Audrey Whaling for their assistance in various parts of the project.

References


Appendix. Competency Testing
All student raters were tested for accuracy in reading the gauges and in administering the protocol according to American Society of Hand Therapists procedures (Fess & Moran, 1981).

Training
The competency training included teaching the theoretical foundation for grip strength testing, demonstrating how to administer the assessment, and showing the students 10 slides of the instrument gauges. The readings for the first 5 slides were discussed as a group. The students recorded their individual reading for the last 5 slides; their readings were then compared with the instructor’s readings. Any discrepancies were resolved with the slides so the student would understand how to accurately read the gauge.

Competency Testing
The instructor squeezed the dynamometer, read it, and recorded the reading in pounds; she then carefully passed it to each of the student raters (up to six), who read it and recorded the reading. This procedure was repeated 9 more times so that each student recorded 10 readings. These readings were then compared with the instructor’s readings, and any errors were discussed. Additional readings were conducted if needed. Each student had to have 10 accurate readings before qualifying as a rater for the study.