Reliability and Validity of the National Institutes of Health Stroke Scale for Neuroscience Nurses

Janice L. Hinkle, RN, PhD, CNRN, FAHA

The National Institutes of Health Stroke Scale (NIHSS) is used widely among stroke neurologists and neuroscience nurses. It is most commonly used with acute ischemic stroke (AIS) but is also used to assess patients after hemorrhagic stroke or a suspected transient ischemic attack. The scale is used widely to assess patient status acutely, to assess treatment efficacy, and to predict outcome. This article provides a review of pertinent information about the reliability, validity, instruction, and competency on the scale for neuroscience nurses.

The NIHSS was introduced as a 15-item scale, later reduced to 14 items developed by stroke neurologists from the University of Cincinnati, the University of Iowa and the National Institutes of Health-National Institute of Neurological Disorders and Stroke.1,2 This graded physical examination is a measure of impairment that assesses level of consciousness using 3 items, speech, language, cognition, inattention, visual field abnormalities, motor strength (4 items), sensory impairment, and ataxia. It was designed specifically for AIS clinical trials, and potential scores range from 0 to 42. A modified version that removed the initial level of consciousness item, as well as the ataxia, facial weakness, and dysarthria items resulted in an 11-item scale with potential scores ranging from 0 to 31.3

Training and certification in use of the NIHSS are recommended.4 Formal training can be found in ≥2 places (www.strokeassociation.com or www.ninds.gov). Since the success of the National Institute of Neurological Disorders and Stroke intravenous recombinant tissue plasminogen activator (rt-PA) trial, it has been acknowledged that performing and documenting a physical examination in the most efficient and useful manner. It is easiest to picture reliability when thinking about physical measures such as weight. For example, when measuring weight, given that all other variables are the same (like how much food was consumed) if a scale weighs a person at 120 pounds today, then that same scale should weigh that person at 120 pounds the next day. Understanding reliability in an observational scale for something like a physical examination is not quite as straightforward.

An important aspect of reliability of the NIHSS is the ability of the measure to produce the same results when used by different individuals. This is called inter-rater or interobserver reliability sometimes reported using an intraclass correlation coefficient. The κ statistic or κ coefficient is also used commonly to indicate the magnitude of agreement between observers. A κ of 1 indicates complete agreement, whereas a κ of 0 indicates that the agreement is because of chance. A κ can be reported as weighted or unweighted.5

An important aspect of inter-rater reliability is agreement on total scores. The overall reliability of the NIHSS in the context of stroke clinical trials is well established.7 One study assessed the reliability of the NIHSS when administered by research nurses.8 Using the NIHSS, 31 consecutive patients who had a stroke were assessed by 2 neurologists and 1 of 2 trained research nurses. There was a high level of agreement for total scores between the 2 neurologists (intraclass correlation coefficient=0.95) and between each neurologist and research nurse (intraclass correlation coefficient=0.92 and 0.96). It was concluded that in both hospital and community settings, trained research nurses can administer the NIHSS with reliability similar to stroke-trained neurologists.8

Another study investigated the reliability of the NIHSS in a representative sample of raters who had completed videotaped certification examinations on the National Stroke association website between 1998 and 2004. There were 7405 unique raters and more than half (54%) of them were registered nurses. Nurses agreed with the most common response on scoring more frequently than physicians (P<0.0001) and 91% if the registered nurses passed the certification examination.9

An additional aspect of inter-rater reliability is agreement on individual item scores within the scale. One study reported moderate to excellent agreement (weighted κ>4) overall in statistical terms reliability is the ability of an instrument to measure something consistently and repeatedly in the same manner. It is easiest to picture reliability when thinking about physical measures such as weight. For example, when measuring weight, given that all other variables are the same (like how much food was consumed) if a scale weighs a person at 120 pounds today, then that same scale should weigh that person at 120 pounds the next day. Understanding reliability in an observational scale for something like a physical examination is not quite as straightforward.

The Reliability

A measurement instrument that is reliable is one that is dependable, stable, consistent, predictable, and accurate across time. In statistical terms reliability is the ability of an instrument to measure something consistently and repeatedly in the same manner. It is easiest to picture reliability when thinking about physical measures such as weight. For example, when measuring weight, given that all other variables are the same (like how much food was consumed) if a scale weighs a person at 120 pounds today, then that same scale should weigh that person at 120 pounds the next day. Understanding reliability in an observational scale for something like a physical examination is not quite as straightforward.

An important aspect of reliability of the NIHSS is the ability of the measure to produce the same results when used by different individuals. This is called inter-rater or interobserver reliability sometimes reported using an intraclass correlation coefficient. The κ statistic or κ coefficient is also used commonly to indicate the magnitude of agreement between observers. A κ of 1 indicates complete agreement, whereas a κ of 0 indicates that the agreement is because of chance. A κ can be reported as weighted or unweighted.5

An important aspect of inter-rater reliability is agreement on total scores. The overall reliability of the NIHSS in the context of stroke clinical trials is well established.7 One study assessed the reliability of the NIHSS when administered by research nurses.8 Using the NIHSS, 31 consecutive patients who had a stroke were assessed by 2 neurologists and 1 of 2 trained research nurses. There was a high level of agreement for total scores between the 2 neurologists (intraclass correlation coefficient=0.95) and between each neurologist and research nurse (intraclass correlation coefficient=0.92 and 0.96). It was concluded that in both hospital and community settings, trained research nurses can administer the NIHSS with reliability similar to stroke-trained neurologists.8

Another study investigated the reliability of the NIHSS in a representative sample of raters who had completed videotaped certification examinations on the National Stroke association website between 1998 and 2004. There were 7405 unique raters and more than half (54%) of them were registered nurses. Nurses agreed with the most common response on scoring more frequently than physicians (P<0.0001) and 91% if the registered nurses passed the certification examination.9

An additional aspect of inter-rater reliability is agreement on individual item scores within the scale. One study reported moderate to excellent agreement (weighted κ>4) overall
between neurologists and research nurses on the majority of the NIHSS items but poor agreement on the limb ataxia item. In another study, a total of 38148 individual NIHSS item responses reported poor agreement on 4 items. In a large sample of clinicians, the agreement on the limb ataxia item was extremely low, whereas the 3 items assessing gaze, aphasia, and facial weakness were low using an unweighted \( \kappa \) statistic. Another study using updated cases reported low agreement on the ataxia and facial weakness items.

### Validity

Validity in behavioral measures refers to how well the instrument actually measures the construct it says it is measuring. For example, if an instrument is measuring disability, is it really measuring disability or is it measuring impairment? The validity of the NIHSS has been studied in several different ways.

The clinical predictive validity of the NIHSS is useful to neuroscience nurses in many ways. There have been a variety of studies in the acute phase of stroke. In a study of 643 patients with AIS, an initial total NIHSS score of \( \geq \) 28 points was predictive of neurological deteriorations within the first week. In the National Institute of Neurological Disorders and Stroke intravenous rt-PA trial patients with an NIHSS, total score of \( >22 \) had a 17% risk of intracranial hemorrhage, whereas patients with a total NIHSS score of \( <10 \) had a 3% risk. A total NIHSS \( >25 \) is an exclusion criterion for rt-PA.

One small study looked at the NIHSS items and found that increases in the loss of consciousness and motor limbs total scores were related to neurological deterioration within the first 120 minutes after administration of intravenous rt-PA. Another study reported that an improvement in the total NIHSS of \( \geq 3 \) points at 15 minutes or of \( >5 \) points at 30 minutes predicts a more favorable outcome and helps identify patients who are not responding to rt-PA. A Swedish study of 347 patients found that for each point on the baseline NIHSS there was an increase in length of stay by 0.8 days and the total length of stay (including rehabilitation) by 3.4 days up to a NIHSS of 19 points. The NIHSS was better at predicting the presence of dysphagia compared with a nursing dysphagia screening tool in a retrospective cohort study of veterans admitted with ischemic stroke.

Several studies have looked at outcome after hospitalization. In a study of 385 patients 3 months after a stroke, a total NIHSS score of \( \geq 15 \) points was associated strongly with the patient being dependent (in a nursing home, chronic home, or substantially dependent on a caregiver). One study reported that patients with a NIHSS total score of \( <10 \) have a more favorable outcome at 1 year compared with patients with an NIHSS score of \( >20 \). In a community-based sample of 377 patients scoring \( <4 \) on baseline NIHSS, 75% were independent 1 year after the stroke, 17% were dependent, and 8% were dead. Predicting outcome makes the NIHSS useful for neuroscience nurses working with families on discharge planning needs.

Factor analysis is a statistical process that is used to establish how individual items cluster around a dimension. Two factors were found in factor analysis of the NIHSS corresponding to left and right brain function in a sample of mild to moderate stroke and another of patients with large strokes. In both studies patients with left brain strokes score 4 points higher on the NIHSS compared with right brain strokes.

### Learning the NIHSS

The NIHSS has been taught in person, with videotapes, DVDs and is now there are professionally filmed cases for training and testing are on the Web. Instruction takes 2 to 3 hours. Still the best method of learning the scale is debated. One study compared the use of videotaped instruction to an interactive computer-assisted instruction to teach the Chinese version of the NIHSS. The nurses with less experience in neurological nursing performed a better assessment after the interactive computer-assisted instruction compared with nurses who were taught with an instructor-led program using the videotape. More research is needed to determine the best method of instruction for the NIHSS.

### Competency

Just because an individual has been educated about the use of a scale and even received a certificate does not indicate they are competent in the performance of the scale. This is particularly true in instances when the scale is not used on a day-to-day basis. A self-assessment of competency of neurological assessment techniques that follows the components of the NIHSS using a clinical skills checklist has been developed. Clearer more work is needed in the area of competency for the NIHSS.

### TAKE-HOME POINTS

- The NIHSS has a high degree of both reliability and validity when used by neuroscience nurses.
- The scale is useful in identifying patients at highest risk for intracranial hemorrhage and those not responding to rt-PA.
- Research is needed to determine the best method of instruction for the NIHSS.
- Further work is needed in the area of competency for neuroscience nurses using the NIHSS.

### Disclosures

None.

### References

1. Brott T, Adams HP Jr, Olinger CP, Marler JR, Barsan WG, Biller J, et al; American Heart Association Council on Cardiovascular Nursing and...

Key Words: reliability and validity
Reliability and Validity of the National Institutes of Health Stroke Scale for Neuroscience Nurses
Janice L. Hinkle

Stroke. 2014;45:e32-e34; originally published online February 4, 2014;
doi: 10.1161/STROKEAHA.113.004243

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/45/3/e32