McGill, R. J. (in press). Test review of NIH Toolbox-Sensation Domain. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The twenty-first mental measurements yearbook* (pp. xx-xx). Buros Center for Testing.

Review of the NIH Toolbox-Sensation Domain, by Ryan J. McGill, Assistant Professor of School Psychology, William & Mary School of Education, Williamsburg, VA:

Description

The NIH Toolbox-Sensation Domain is a core component of the National Institutes of Health Toolbox for Assessment of Neurological and Behavioral Function—a multidimensional set of brief tests designed to assess cognitive, sensory, motor, and emotional functioning that can be used by researchers and clinicians in a variety of settings to measure outcomes in clinical trials and longitudinal or epidemiological studies (National Institutes of Health [NIH] & Northwestern University, 2006-2018). The battery has been validated across the lifespan and norms have been developed for various measures for participants ages 3 to 85 years and can be administered in English or Spanish. In the Toolbox Brochure, it is noted that use of measures from the NIH Toolkit ensures that "assessment methods and results can be used for comparisons across existing and future studies" (National Institutes of Health & Northwestern University, 2017, p. 2) and that the Toolbox functions as a sort of "common currency" for neuroscience research.

Sensation is defined in the Administrator's Manual as the "biochemical and neurologic process of detecting incoming impulses as nervous system activity" (National Institutes of Health & Northwestern University, 2006-2018). The Sensation Domain contains six measures (Words-in-Noise Test, Visual Acuity Test, Odor Identification Test, Pain Intensity Survey, Pain Interference Survey, and Regional Taste Intensity Test) that assess functioning in Audition, Vision, Olfaction, Perceived Pain, and Taste respectively. Each measure is delivered exclusively through the NIH Toolbox iPad® app. In order to use NIH Toolbox measures, users must

download the app from the APP Store and purchase a 12-month subscription ($499) that provides access to all measures in the Toolbox. It should be noted that several Sensation measures require users to purchase additional iPad accessories and assessment materials (~$1,000) that are not covered in the cost of the yearly subscription. For example, in order to administer the Regional Taste Intensity Test, clinicians must obtain the necessary laboratory equipment to produce standardized Quinine and Sodium Chloride solutions that are administered to the tip of an examinees' tongue as well as whole mouth during the administration of that test. Detailed lists of required Toolbox equipment and materials are provided in appendices in the Administrator's Manual which is located at the NIH Toolbox website (www.nihtoolbox.org).

Development

The genesis of the NIH Toolbox can be traced to 2004 when the NIH formed a coalition called the Blueprint for Neuroscience Research. The goal of the coalition was to develop a set of new standardized assessment tools that could be used by researchers across the country to facilitate discoveries in neuroscience research. In 2006, a large-scale grant was awarded to a team of 250 scientists from nearly 80 academic institutions to develop Toolbox measures. Many of the Sensation measures were adapted from existing technologies or surveys (e.g. San Diego Odor Identification Test, Patient-Reported Outcomes Measurement Information System [PROMIS®] Pain Interference Survey). Each measure was uniquely calibrated and formatted for standardized digital administration using item response theory and computer adaptive testing procedures. This process allows users to assess various Sensation abilities in a relatively short duration of time without sacrificing psychometric integrity.

The NIH Toolbox was officially released in 2011 and designed to be administered through a web-based data collection platform. In 2015 an iPad app was developed to supplant the web platform and access to the online assessment center was eventually closed. It should be

noted that the Sensation Domain originally contained additional Vestibular, Auditory, and Tactile measures and validation evidence for these indicators has been reported (Rine et al., 2013), However, these measures are not presently available to users in the iPad app and are not reviewed here. During the migration to the iPad app, internal validation studies were conducted to establish the equivalency of Sensation Domain scores across platforms. Results of those studies indicate that score corrections are not needed for that domain. Although Toolbox measures have undergone a continuous validation process since their development, inspection of test Technical Manuals reveal that no significant modifications have been made to any of Sensation Domain measures since their initial validation.

The age ranges for Sensation Domain measures vary by test. Whereas performance measures can generally be administered from ages 3-85, the Pain surveys can only be administered to examinees ages 18 and older. All six measures can be administered in approximately 30 minutes and most of the tests can be delivered with the iPad placed directly in front of the examinee with the assessor seated next to them. Verbal directions for each test are provided on the iPad screen and modifications are available for younger examinees. However, several measures require additional time to prepare assessment materials and iPad accessories (e.g., Regional Taste Intensity Test, Visual Acuity Test). With the exception of the Pain measures, all of the tests require users to identify and respond to various sensory stimuli. For example, identifying words spoken through earphones amidst increasing levels of background noise and letters that gradually decrease in size on a conventional visual array. A *Scoring and Interpretation Guide for the iPad* (National Institutes of Health & Northwestern University, 2016), Technical Manuals for all of the Sensation tests, Administrator's Manual, and a comprehensive eLearning module can be accessed at the Toolbox website.

Technical

Although Technical Manuals are provided online for each test in the Sensation Domain, conventional reliability and validity information are not disclosed in these documents. Instead, readers are directed to a panoply of research articles reporting results from validation studies across the age ranges of the tests typically including 400-500 participants. This reviewer was provided with detailed list of Sensation Domain articles by the test publisher and the information provided below was extracted from the articles from that list (These references are also available at the Toolbox website).

Standardization

The Technical Manuals for each Sensation measure briefly describe the NIH Toolbox total standardization sample and users are directed to Beaumont et al. (2013) for more details about the sampling plan. The standardization sample contains 4,859 English and Spanish speaking participants, ages 3-85, and it is reported that the sample was representative of the United States population based on key demographic variables from the 2010 United States Census Survey. Whereas Beaumont and colleagues (2013) outline numerous desired characteristics for the normative sample that are consistent with best practice standards, the actual obtained results on many key demographic variables are not reported in available technical documentation so users are unable to independently evaluate the degree to which obtained estimates match many of the intended parameters.

Primary Toolbox measures were administered to participants at 10 sites across the United States. Separate norms are provided for males and females (disaggregated by primary language) for 21 age brackets (each year from ages 3-17, 18-29, 30-39, 40-49, 50-59, 60-69, and 70-85). The raw scores for each test are tabulated automatically by the app during the course of administration and a score report is generated at the conclusion of the assessment with raw scores transformed into various normative scores. For performance measures, age-corrected

standard scores which compare an examinees' performance to other examinees their age (M = 100, SD = 15), uncorrected standard scores which compare an examinees' performance to the total normative sample (M = 100, SD = 15), fully corrected standard scores which correct for age and other key demographic variables (M = 50, SD = 10), and percentile ranks are available. For person-reported outcome (PRO) measures, such as the Pain surveys, uncorrected T-scores (M = 50, SD = 10) are reported. Two of the Sensation measures report scores that are idiosyncratic to those tests. For the Words-In-Noise Test, the obtained raw score for each ear is converted to a decibels of signal-to-noise ratio (dB S/N) score ranging from -2.0 to 26.0 with lower scores indicative of better performance. Additionally, for the Visual Acuity Test, the raw score is transformed logarithmically into a conventional Snelling Visual Acuity score. In terms of clinical interpretation, practitioners are encouraged to interpret most of the Toolbox scores primarily as screening indicators. For example, in the Scoring and Interpretation Guide, it is recommended that examinees who obtain a Snelling equivalent score of 20/40 or worse should be referred to a eye care professional for additional testing (National Institutes of Health & Northwestern University, 2016, p. 24).

Although descriptive statistics, disaggregated by age, are provided for each normative age bracket in respective test Technical Manuals, raw score to normative score conversions are not available. Therefore this reviewer is unable to evaluate whether the item gradients at each age are sufficient. Given the relatively small number of items in each of the Sensation measures, it raises concern about potential item density issues within the scales. According to Wasserman and Bracken (2013), inadequate item density across the distribution of a variable reduces the sensitivity and discrimination of a test. No normative information are reported for either of the shortened Pain surveys.

Reliability

Within the reference list provided, three studies were found to report conventional reliability evidence. Test-retest reliability coefficients for the Odor Identification Test reported by Dalton et al. (2013) ranged from .45-.57 which are considered low. In contrast, Rine and colleagues (2012) report strong test-retest estimates (.84-.91) for the Visual Acuity Test in the form of intra-class correlation coefficients (ICCs). Finally, Mennella, Lukasewycz, Griffith, and Beauchamp (2011) report low to moderate (.42-.65) internal consistency estimates also in the form of ICCs. It should be noted that the use of ICCs with these data may be problematic given the skew that is likely in the underlying distributions. As noted by Bobak, Barr, and O'Malley (2018), the ICC coefficient assumes that data are normally distributed and when this assumption is violated reliability estimates may be inflated.

Validity

Similar to reliability, validity evidence for the Sensation Domain varies by test. Whereas content validity for the measures was established largely via a series of articles published in a special issue of *Neurology* in 2013, compelling structural validity evidence was unable to be located. To date, construct validity evidence is largely limited to concurrent validity in the form of correlations with existing measures for the Visual Acuity Test and Regional Taste Intensity Test and evidence of developmental differences for the former as well as the Odor Identification Test. Available diagnostic utility studies have been mixed. In a comprehensive study of Toolbox measures with neurologic patients, Carlozzi et al. (2017) found that individuals with Traumatic Brain Injury where at-risk for olfactory dysfunction. However, Abasaeed and colleagues (2018) found that scores from the Regional Taste Intensity Test did not differ among patients receiving stem cell transplantation over the course of treatment, raising concern about the sensitivity of the measure.

Commentary

As with any test, the NIH-Sensation Domain has strengths and weaknesses. The iPad app provides an innovative platform for clinical assessment and, in terms of assessment, has few flaws. Most of the measures require minimal training or advanced setup although users are encouraged to supplement the Administrator's Manual with available video demonstrations to facilitate assessment fidelity. Clinicians unfamiliar with basic laboratory techniques will likely need additional practice preparing assessment materials for the Regional Taste Intensity Test. Many of the embedded links in the app to technical documentation on the Toolbox website are broken which makes it difficult to access those materials. Thus, users will likely have to consult the website via a conventional computer in order to fully ascertain the psychometric properties of the instrument.

In spite of the conventional appeal of the measures, reliability and validity for the tests, in their present form, remains underdeveloped. Whereas extensive validation work has been conducted on other Toolbox domains, the validation network for the Sensation Domain measures presently consists of a small series of articles mostly using preliminary versions of the tests. In some cases, stimuli and assessment methods differ significantly from current versions of the measures (e.g., Smutzer, Desai, Coldwell, & Griffith, 2013). As a result, it is unclear whether this information generalizes to the Sensation tests presently available in the iPad app. Even so, it is believed that the development of the current delivery system will be instrumental for furthering existing validation efforts and it is anticipated that additional research on the measures should follow. As that information accumulates, users would benefit from a reformatting of the Technical Manuals to include this information rather than directing readers to research articles which may be difficult for clinicians to locate.

Summary

Overall, the NIH-Sensation Domain has many strengths. For the most part, the app is an excellent technological innovation and is of exceptional quality. Whereas the technical documentation for many of the measures remains underdeveloped, it is more than adequate to support their use in research settings, which is the fundamental goal underlying the development of the NIH Toolbox. However, clinicians looking to replace existing instrumentation in clinical practice settings, should employ Sensation Domain measures with caution until additional validity evidence is presented, in particular, diagnostic validity evidence, indicating that the measures have adequate sensitivity and specificity to identify targeted pathologies.

References

Abasaeed, R., Coldwell, S. E., Lloid, M. E., Soliman, S. H., Macris, P. C., & Schubert, M. M. (2018). Chemosensory changes and quality of life in patients undergoing hematopoietic stem cell transplantation. *Supportive Care in Cancer, 26,* 3553-3561. doi: 10.1007/s00520-018-4200-7

Beaumont, J. L., Havlik, R., Cook, K. F., Hays, R. D., Wallner-Allen, K., Korper, S. P., . . . Gershon, R. C. (2013). Norming plans for the NIH Toolbox. *Neurology, 80 (11 Supplement 3),* S87-S92. doi: 10.1212/WNL.0b013e3182872e70

Bobak, C. A., Barr, P. J., & O'Malley, A. J. (2018). Estimation of an intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. *BMC Medical Research Methodology, 18,* 93. doi: 10.1186/s12874-018-0550-6

Carlozzi, N. E., Goodnight, S., Casaletto, K. B., Goldsmith, A., Heaton, R. K., Wong, A. W. K., . . . Tulsky, D. S. (2017). Validation of the NIH Toolbox in individuals with neurologic disorders. *Archives of Clinical Neuropsychology, 32,* 555-573. doi: 10.1093/arclin/acx020

Dalton, P., Mennella, J. A., Doty, R. L., Murphy, C., Frank, R., Hoffman, H. J., . . . Slotkin, J. (2013). Olfactory assessment using the NIH Toolbox. *Neurology, 80(11 Supplement 3),* S32-S36 doi: 10.1212/WNL.0b013e3182872eb4

Mennella, J. A., Lukasewycz, L. D., Griffith, J. W., & Beauchamp, G. K. (2011). Evaluation of the Monell Forced-Choice, paired-comparison tracking procedure for determining sweet

taste preferences across the lifespan. *Chemical Senses, 36,* 345-355. doi: 10.1093/chemse/bjq134

National Institutes of Health, & Northwestern University (2006-2018). *NIH Toolbox® for Assessment of Neurological and Behavioral Function administrator's manual*. Location: Author.

National Institutes of Health, & Northwestern University (2016). *NIH Toolbox® scoring and interpretation guide for the iPad*. Location: Author.

National Institutes of Health, & Northwestern University (2017). *NIH Toolbox® for Assessment of Neurological and Behavioral Function brochure*. Location: Author.

Rine, R., Roberts, D., Corbin, B. A., McKean-Cowdin, R., Varma, R., Beaumont, J., . . . Schubert, C. (2012). A new portable tool to screen vestibular and visual function in children and adults—NIH Toolbox. *Journal of Rehabilitation Research and Development, 49*, 209-220. doi: 10.1682/JRRD.2010.12.0239

Rine, R. M., Schubert, M. C., Whitney, S. L., Roberts, D., Redfern, M. S., . . . Slotkin, J. (2013). Vestibular function assessment using the NIH Toolbox. *Neurology, 80 (11 Supplement 3),* S25-31. doi: 10.1212/WNL.0b013e318287c6a

Smutzer, G., Desai, H., Coldwell, S. E., & Griffith, J. W. (2013). Validation of edible taste strips for assessing PROP taste perception. *Chemical Senses, 38*, 529-539 doi: 10.1093/chemse/bjt023

Waserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.). *Handbook of psychology: Assessment psychology* (2nd ed., Vol. 10, pp. 50-81). Hoboken, NJ: John Wiley.