NIH Toolbox®

Scoring and Interpretation Guide

for the iPad

Updated 5/25/2021

NIH Toolbox® Scoring and Interpretation Guide for the iPad

Table of Contents

Introduction

- The *NIH Toolbox® Scoring and Interpretation Guide for the iPad* is designed to be a resource to NIH Toolbox users. It provides basic information about how each of the measures that comprise the NIH Toolbox is scored, as well as the potential uses and meaning of the available scores. It is not intended to be a highly technical document statistically, though some technical details are included as necessary to aid understanding. The goal of this guide is to provide information on how each measure is scored – even though this scoring is done automatically by the available NIH Toolbox software – so that users will have an essential understanding of the numbers they see on a score output report or in a data file and how they were created.

This guide is organized to provide:

1) Overarching information about the NIH Toolbox scoring approach and the scores generally available for each test, and

2) Specific information on each measure and its scoring and interpretation, by domain (Cognition, Motor, Sensation, Emotion).

For those readers who want to learn about specific measures rather than the entire NIH Toolbox, we recommend that you read the section, "Scores Available in NIH Toolbox," and then go to the domain section(s) of interest.

A greater amount of interpretive information is provided for some measures than for others, primarily because more can comfortably be asserted about the interpretation of those measures; throughout this guide, we have attempted to avoid making any interpretive statements that exceed the limits of the available data or scientific knowledge in that area. Rather, the intent of the brief interpretation statements is to provide some basic information to researchers who may be less familiar with a given measure or domain, and who would thus benefit from some guidance about appropriate limits to interpretation and/or specific score ranges that may merit individual follow-up.

Scores Available in NIH Toolbox®

General Scoring Approach

NIH Toolbox scoring is intended to provide a range of useful and meaningful information to support those with all levels of expertise. Scores offer a snapshot of individuals' and groups' levels of functioning at a given point. We are providing two different normative scores in keeping with professional standards in the given fields: one set for performance measures, the other score for person-reported-outcome (PRO) measures. In addition, all normative scores were derived separately for English and Spanish speakers. The normative scores offer the same types of information, with each using the labels commonly found on such measures (PRO vs. performance types).

Performance measures have three types of scores: Age-Corrected Standard Scores, for which the normative mean is 100 and the standard deviation (SD) is 15 (commonly referred to as Standard Scores); Uncorrected Standard Scores (mean = 100, SD = 15); and Fully Corrected Scores, which are primarily intended for neuropsychological applications and correct for age and other demographic characteristics (education, sex, and race/ethnicity) that may affect the performance of normal people. To distinguish them from Age-Corrected Standard Scores, these Fully Corrected Scores are based upon a T-Score metric, with a normative mean of 50 and an SD of 10. Following convention in the field, PRO scores are also on this T-Score metric; for adults, these have no demographic correction and reflect census-based averages for the U.S. adult population, whereas (again, following convention) children's T-Scores are corrected for age and sex.

More information about how to understand and interpret scores is provided in each measure-specific (or domain-specific) section of this guide. Also, while normative scores are provided for most NIH Toolbox measures, certain exceptions exist; these are noted in each measure-specific section. Although, as noted, performance and PRO measure normative scores are similar in nature, they are described separately below for ease of understanding.

Normative scores provided for performance measures

- **Age-Corrected Standard Score:** This score compares the score of the test-taker to those in the NIH Toolbox nationally representative normative sample at the same age, where a score of 100 indicates performance that was at the national average for the test-taking participant's age. Age-corrected standard scores were derived separately for children (ages 3-17) and adults (ages 18-85). A score of 115 or 85, for example, would indicate that the participant's performance is 1 SD above or below the national average, respectively, when compared with like-aged participants. Higher scores indicate better performance.
- **Fully Corrected T-Score:** This score (which has a mean of 50 and an SD of 10, unlike age-corrected performance normative scores) compares the score of the test-taker to those in the NIH Toolbox nationally representative normative sample, while adjusting for key demographic variables collected during the NIH Toolbox national norming study. These variables, which include age, gender, race/ethnicity (white/Asian, black, Hispanic, multiracial), and educational attainment (for ages 3-17, parent's education is used; education is often used as a proxy for socioeconomic status), are often found to impact performance within a given domain, and thus a "fully corrected" score is provided that allows for comparison within a narrower grouping. For example, a fully corrected T-score of 50 on the NIH Toolbox Grip Strength Test for a 35-year-old, Hispanic, white male with a college education indicates performance that was at the national average for *Hispanic, white males with*

college education, age 35. This can be useful information since, in this example, notable differences in grip strength are typical for males vs. females; using the fully corrected score allows the user to evaluate male (or female) performance relative to normal people not only of the same sex, but with the same overall demographic characteristics (as much as possible, normal people just like him/her). Higher fully corrected T-scores indicate better performance. Users wishing to convert this T-Score to a Standard Score or percentile can use the chart in Appendix A.

- **Uncorrected Standard Score:** This is an additional available score, which also uses a standard score metric (normative mean = 100, SD = 15). It compares the performance of the test-taker to those in the entire NIH Toolbox nationally representative normative sample, regardless of age or any other variable. The Uncorrected Standard Score provides a glimpse of the given participant's overall performance or ability when compared with the general U.S. population. This score may be most useful when trying to gauge one's overall level of functioning, not in the context of age, gender or other demographic factors. It may also be of interest when monitoring performance over time. Higher Uncorrected Standard scores indicate better performance.
- **Percentiles:** A Percentile represents the percentage of people nationally *above whom* the participant's score ranks (the comparison group will be based on whichever normative score is used). It is simply a transformation of the participant's normative score (Age-Corrected Standard Score, Fully Corrected T-Score, or Uncorrected Standard Score) into a format that many consider more easily understood. For example, if a 14-year-old attains a national percentile of 84 on a given Toolbox performance measure, it means that he/she performed better than 84 percent of 14-year-olds in the large Toolbox census-weighted (for sex, education, and race/ethnicity) national norming study. More generally, it suggests that this 14-year-old performs better than 84 percent of 14-year-olds *in the general population.* For ease of understanding, this percentile corresponds to an Age-Corrected Standard Score of 115 – exactly 1 SD above the mean of 100. While percentiles are not provided in the scoring output file, each can be readily looked up for any Standard Score or T-Score by using the simple conversion chart provided in Appendix A at the end of this guide.

Normative scores provided for all PRO measures

- **Uncorrected T-Score:** This score, provided for participants of all ages on PRO measures, compares the performance of the test-taker to those in the entire NIH Toolbox nationally representative normative pediatric or adult (as appropriate) sample, regardless of age or any other variable. The Uncorrected Standard Score provides a glimpse of the given participant's overall performance when compared with the general U.S. population. For all ages, an uncorrected T-score is provided. The uncorrected score may be most useful when trying to gauge one's overall level of functioning, not in the context of age, gender, or other demographic factors. It may also be of interest when monitoring performance over time. It should be noted that while previous versions of the NIH Toolbox have included more normative score types for PRO measures, new normative scoring approaches do not support such distinctions by demographic variables, and they were, therefore, not included in the iPad app version of NIH Toolbox.
- **Age- and Gender-Corrected T-Score for Children (ages 3-17):** These scores are only provided for children and only for the NIH Toolbox Emotion measures. For these scores, corrections are made both for age and for sex (they are the factors that can lead to significantly and meaningfully different scores for these ages, based on analyses of the NIH Toolbox normative study data). There are two main reasons for providing age- and gender-corrected PRO scores for children on the NIH Toolbox Emotion battery: 1) somewhat different instruments are used for different ages, including that they are based only on parent report for ages 3-7, both parent report and self-report for ages 8-

12, and only self-report for ages 13-17; and 2) it is generally considered not appropriate or desirable to use the same normative standards and expectations for boys and girls at different ages (as an extreme example, for a 3-year-old boy and a 17-year-old girl).

NIH Toolbox® Cognition Domain

The Cognition Domain measures several aspects of cognitive functioning for ages 3-85, including language, episodic memory, executive function, working memory and processing speed. Specific recommended age bands for each measure are noted below; available cognition batteries and composite scores are described at the end of this section. Seven core NIH Toolbox Cognition measures, two supplemental measures (variants of existing tests for use with young children), and four NIH Toolbox composite scores are available.

Cognition Core Measures

NIH Toolbox Picture Vocabulary Test (TPVT)

Description: This measure of receptive vocabulary is administered in a computerized adaptive format. That is, the next question a participant receives depends on his/her response to the previous question; Computer Adaptive Testing (CAT) ensures a test that is tailored to the participant's needs. The respondent is presented with an audio recording of a word and four photographic images on the iPad screen, and is asked to select the picture that most closely matches the meaning of the word. This test takes approximately four minutes to administer and is recommended for ages 3-85. Separate but parallel vocabulary tests have been developed in English and Spanish.

Scoring Process: Item Response Theory (IRT) is used to score the TPVT. A score known as a *theta score* is calculated for each participant; it represents the relative overall ability or performance of the participant. A theta score is very similar to a z-score, which is a statistic with a mean of zero and a standard deviation of one. Age-Corrected and Uncorrected Standard Scores, and Fully Corrected T-Scores are provided for the TPVT, and associated Percentiles can be found in Appendix A.

Interpretation: The TPVT is a measure of general vocabulary knowledge and is considered to be a strong measure of crystallized abilities (those abilities that are more dependent upon past learning experiences and are relatively consistent across the adult life span). To interpret individual performance, one can evaluate all three types of scores. A participant's Age-Corrected Standard score at or near 100 indicates vocabulary ability that is average for the age level. Scores around 115 suggest above-average vocabulary ability, while scores around 130 suggest superior ability – in the top 2 percent nationally for age, based on NIH Toolbox normative data. Conversely, a score of 85 suggests below-average vocabulary ability, while a score in the range of 70 or below suggests markedly low language ability (bottom 2 percent nationally), which also is likely to be associated with difficulties in school (for children) or trouble functioning in work environments with a language demand.

An Uncorrected Standard score allows us to view the participant's performance in comparison to the census-matched U.S. population, allowing for a more absolute view of the participant's ability and allows for gauging true improvement or decline from previous assessments. The Fully Corrected T-Scores have been statistically adjusted to level the playing field interpretively, such that an individual's score can be compared to a narrower group, more similar demographically. It should be noted that a raw score does not provide relevant information on a computer-adaptive test. (Raw scores are useful for monitoring absolute improvement/decline over time when statistical transformations are not used in the scoring process, such as occur in IRT-based scoring or in the Flanker or DCCS measures, described below.) Thus, an increase in the Uncorrected Standard score (or the participant's obtained theta value, alternatively) represents real improvement by the participant in vocabulary knowledge; however, this individual's Age-Corrected Standard Score may or may not have increased, depending on how his/her

performance at Time 1 and Time 2 compared to the age cohorts used in the national norms. An individual who has made small gains in overall knowledge may still have regressed when compared with age-similar peers if the national sample of peers made larger gains in knowledge over the same period. Thus, one can see the value of the variety of NIH Toolbox scores provided. Each language version of the TPVT is calibrated independently using language-specific items administered to a language-specific cohort. Therefore, TPVT computed scores are not compatible between different languages (similar scores on the English and Spanish picture vocabulary tests are not comparable).

NIH Toolbox Oral Reading Recognition Test (Reading)

Description: Separate but parallel reading tests have been developed in English and Spanish. In either language, the participant is asked to read and pronounce letters and words as accurately as possible. The test administrator scores them as right or wrong. For the youngest children, the initial items may require them to identify letters (vis-à-vis non-letter symbols) and to identify a specific letter in an array of four symbols. The test is given via CAT and requires approximately three minutes. This test is recommended for ages 7-85, but is available for use as young as age 3, if desired. Normative scores are available for ages 3-6, but the test may not be appropriate for all children, especially those who cannot yet identify any letters of the alphabet.

Scoring Process: IRT is used to score the Reading Test. A theta score is calculated for each participant, representing the overall reading ability or performance of the participant. All NIH Toolbox normative scores are provided for the Reading Test.

Interpretation: The Reading Test is a measure of reading decoding skill and, like vocabulary, is considered among the crystallized abilities; those abilities are generally more dependent upon past learning experiences and consistent across the adult life span. To interpret individual performance, one can evaluate all three types of standard scores plus the percentiles (see Appendix A); higher scores indicate better reading ability within the normative standard being applied. The Reading Uncorrected Standard score (or the theta score, alternatively) can also be useful in evaluating pure change in performance from one assessment to another. For example, a higher Uncorrected Standard score (or theta score) for Reading would mean that the participant is able to correctly identify more difficult words on the subsequent assessment, which may indicate developmental growth or a return to a previous higher level of functioning. Each language version of the Reading test is calibrated independently using language-specific items administered to a language-specific cohort. Therefore, Reading computed scores are not compatible between different languages (similar scores on the English and Spanish reading tests are not comparable). Note further that reading decoding skill is generally considered easier in the Spanish language than it is in English due to the presence of fewer linguistic exceptions.

NIH Toolbox Flanker Inhibitory Control and Attention Test (Flanker)

Description: The Flanker task measures both a participant's attention and inhibitory control. The test requires the participant to focus on a given stimulus while inhibiting attention to stimuli (fish for ages 3-7 or arrows for ages 8-85) flanking it. Sometimes the middle stimulus is pointing in the same direction as the "flankers" (congruent) and sometimes in the opposite direction (incongruent). Twenty trials are conducted for ages 8-85; for ages 3-7, if a participant scores ≥ 90% on the fish stimuli (with no more than one congruent and one incongruent trial incorrect), 20 additional trials with arrows are presented. The test takes approximately three minutes to administer. This test is recommended for ages 3-85.

Scoring Process: Scoring is based on a combination of accuracy and reaction time and is identical for both the Flanker and DCCS measures (described below). A 2-vector scoring method is employed that uses accuracy and reaction time, where each of these "vectors" ranges in value between 0 and 5, and the computed score, combining each vector score, ranges in value from 0-10. For any given individual, accuracy is considered first. If accuracy levels for the participant are less than or equal to 80%, the final "total" computed score is equal to the accuracy score. If accuracy levels for the participant reach more than 80%, the reaction time score and accuracy score are combined.

Accuracy Vector

There are 40 possible accuracy points:

- Flanker
  - Fish: 20 Points
  - Arrows: 20 Points

Individuals age eight and older automatically receive 20 accuracy points for the Fish Trials of the Flanker. (It was determined previously that they typically score at the ceiling on these trials.) These "free" points are *not* reflected in the raw score in the score report, which only counts the number of administered items with a correct response; however, they are included in the computed score calculation.

The accuracy score varies from 0 to 5 points. For every correct behavioral response, a participant receives a value of 0.125 (5 points divided by 40 trials) added to his/her score for Flanker:

Flanker Accuracy Score = 0.125 * Number of Correct Responses

Reaction Time Vector

The task-specific reaction time scores are generated using individuals' raw, incongruent median reaction time score from the Flanker. Median reaction time values are computed using only correct trials with reaction times greater than or equal to 100 ms and reaction times no larger than 3 SDs away from the individual's mean (for respective trial type). **Children ages 3-7 do not get reaction time scores unless they proceed to the "arrow" portion of the test, based on high performance on the test's "fish" portion.**

Like the accuracy score, the reaction time score ranges from 0 to 5 points. One issue regarding reaction time data is that it tends to have a positively skewed distribution. A log (Base 10) transformation is therefore applied to each participant's median reaction time score from the Flanker, creating a more normal distribution of scores. Based on data from a validation study of the NIH Toolbox Cognition Battery, the minimum median reaction time for scoring is set to 500 ms and the maximum reaction time for scoring is 3,000 ms. Participants with median reaction times that fall outside this range but within the allowable range of 100 ms – 10,000 ms are truncated (i.e., reaction times between 3,000 ms and 10,000 ms are set equal to 3,000 ms) for the purpose of score calculation. Scoring of the validation data indicates that this truncation does not introduce any problems with regard to ceiling or floor effects. Log values are algebraically rescaled from a log(500)-log(3,000) range to a 0-5 range. Note that the rescaled reaction time scores are reversed; smaller reaction time log values are at the upper end of the 0-5 range while larger log values are at the lower end of the range. The formula for rescaling is:

$$\text{Reaction Time Score} = 5 - \left| 5 * \left| \frac{\lceil \log RT - \log(500) \rceil}{\lfloor \log(3000) - \log(500) \rfloor} \right| \right|$$

Once these reaction time scores are obtained, they are added to the accuracy scores for participants who achieved the accuracy criterion of better than 80%. For participants who fail to reach this criterion, only accuracy scores are used. This combination score is then converted to normative scores.

Interpretation: The Flanker is a measure of executive function, specifically tapping inhibitory control and attention. It is considered a "fluid ability" – the capacity for new learning and information processing in novel situations – measure, in which performance reaches a peak in early adulthood, then tends to decline across the life span (based on health and individual factors, of course). To interpret individual performance, one can evaluate all three types of normative scores, in which higher scores indicate higher levels of ability to attend to relevant stimuli and inhibit attention from irrelevant stimuli. In addition to the three normative scores provided, the Flanker Computed score provides a way of gauging raw improvement or decline from Time 1 to Time 2 (or subsequent assessments). This computed score ranges from 0-10, but if the score is less than 4, it indicates that the participant did not score high enough in accuracy (80 percent correct or less) to receive a reaction time score. A change in the participant's score from Time 1 to Time 2 represents absolute change in the level of performance for that individual since the previous assessment. One can also put such a score in a different context by comparing normative scores from Time 1 to Time 2, which will show the participant's performance relative to others (specific comparisons depending on which score is used).

NOTE: A new "experimental" version of Flanker "with Developmental Extension" for ages 3-7 is now also available for researchers wishing to more thoroughly evaluate those who are not able to perform adequately on the standard version of Flanker described here. This experimental version starts with the standard version of Flanker, but drops down to "developmental extension" items for those who cannot complete practice items successfully or who otherwise perform below chance levels on standard test items. All scores described above are also available for this experimental version, as well as raw scores for the experimental items. This version of the test is likely to take longer to administer and is only recommended in specific research cases (to be defined by the investigator) at this time.

NIH Toolbox Dimensional Change Card Sort Test (DCCS)

Description: DCCS is a measure of cognitive flexibility. Two target pictures are presented that vary along two dimensions (e.g., shape and color). Participants are asked to match a series of bivalent test pictures (e.g., yellow balls and blue trucks) to the target pictures, first according to one dimension (e.g., color) and then, after a number of trials, according to the other dimension (e.g., shape). "Switch" trials are also employed, in which the participant must change the dimension being matched. For example, after four straight trials matching on shape, the participant may be asked to match on color on the next trial and then go back to shape, thus requiring the cognitive flexibility to quickly choose the correct stimulus. This test takes approximately four minutes to administer and is recommended for ages 3-85.

Scoring Process: Scoring is based on a combination of accuracy and reaction time. A 2-vector scoring method is employed that uses accuracy and reaction time, where each of these "vectors" ranges in value between 0 and 5, and the computed score, combining each vector score, ranges in value from 0-10. For any given individual, accuracy is considered first. If accuracy levels for the participant are less than or equal to 80%, the final "total" computed score is equal to the accuracy score. If accuracy levels for the participant reach more than 80%, the reaction time score and accuracy score are combined.

Accuracy Vector

There are 40 possible accuracy points:

- DCCS
  - Pre-Switch (before changing to the other dimension): 5 Points
  - Post-Switch: 5 Points
  - Mixed Trials: 30 Points

Individuals age 8 and older automatically receive 10 accuracy points for the Pre-Switch and Post-Switch trials of the DCCS. These "free" points are *not* reflected in the raw score in the score report, which only counts the number of administered items with a correct response; however, they are included in the computed score calculation.

The accuracy score will vary from 0 to 5 points. For every correct behavioral response, a participant receives a value of 0.125 (5 points divided by 40 trials) added to his/her score for DCCS:

DCCS Accuracy Score = 0.125 * Number of Correct Responses

Reaction Time Vector

The task-specific reaction time scores are generated using individuals' raw, non-dominant dimension (the dimension cued less frequently for sorting during the mixed trials) median reaction time score from the DCCS. Median reaction time values are computed using only correct trials with reaction times greater than or equal to 100 ms and reaction times no larger than 3 SDs away from the individual's mean (for respective trial type). Children ages 3-7 do not get reaction time scores unless they proceed to the "mixed trials" portion of the test, based on high performance on the test's "pre-switch" and "post-switch" portions.

Like the accuracy score, the reaction time score ranges from 0 to 5 points. Reaction time data tend to have a positively skewed distribution. A log (Base 10) transformation is therefore applied to each participant's median reaction time score from the DCCS and Flanker, creating a more normal distribution of scores. Based on the validation data, the minimum median reaction time for scoring is set to 500 ms and the maximum reaction time for scoring is 3,000 ms. Participants with median reaction times that fall outside this range but within the allowable range of 100 ms – 10,000 ms will be truncated (i.e., reaction times between 3,000 ms and 10,000 ms will be set equal to 3,000 ms) for the purpose of score calculation. Scoring of the validation data does not indicate that this truncation introduces any problems with regard to ceiling or floor effects. Log values will be algebraically rescaled from a log(500)-log(3,000) range to a 0-5 range. Note that the rescaled reaction time scores will be reversed; smaller reaction time log values will be at the upper end of the 0-5 range while larger log values will be at the lower end of the range. The formula for rescaling is:

$$\text{Reaction Time Score} = 5 - \left| 5 * \left[\frac{\log RT - \log(500)}{\log(3000) - \log(500)} \right] \right|$$

Once these reaction time scores are obtained, they are added to the accuracy scores for participants who achieved the accuracy criterion of better than 80%. For participants who fail to reach this criterion, only accuracy scores are used. This combination score is then converted to normative scores.

Interpretation: The DCCS is a measure of executive function, specifically tapping cognitive flexibility. It is considered a "fluid ability" measure, like Flanker, with performance generally increasing through childhood and then declining across the adult age span. To interpret individual performance, one can evaluate all three types of normative scores, where higher scores indicate higher levels of cognitive

flexibility. In addition to the three normative scores provided, the DCCS Computed score provides a way of gauging raw improvement or decline from Time 1 to Time 2 (or subsequent assessments). This computed score ranges from 0-10, but if the score is less than 4, it indicates that the participant did not score high enough in accuracy (80 percent correct or less) to receive a reaction time score. A change in the participant's score from Time 1 to Time 2 represents an absolute change in the level of performance for that individual since the previous assessment. One can also put such a score in a different context by comparing normative scores from Time 1 to Time 2, which will show the participant's performance relative to others (specific comparisons depending on which type of Standard or T-score is used).

NOTE: A new "experimental" version of DCCS "with Developmental Extension" for ages 3-7 is now also available for researchers wishing to more thoroughly evaluate those who are not able to perform adequately on the standard version of DCCS described here. This experimental version starts with the standard version of DCCS, but drops down to "developmental extension" items for those who cannot complete practice items successfully or who do not perform adequately on subsequent sections of the test. The "developmental extension" items are "scaffolded" such that they start at a level of difficulty that is based on which standard DCCS items the participant did not complete successfully. All scores described above are also available for this experimental version, as well as raw scores for the experimental items. This version of the test is likely to take longer to administer and is only recommended in specific research cases (to be defined by the investigator) at this time.

NIH Toolbox Picture Sequence Memory Test (PSMT)

Description: The Picture Sequence Memory Test is a measure developed for the assessment of episodic memory for ages 3-85 years. It involves recalling increasingly lengthy series of illustrated objects and activities that are presented in a particular order on the iPad screen, with corresponding audio-recorded phrases played. The participants are asked to recall the sequence of pictures demonstrated over two learning trials; sequence length varies from 6-18 pictures, depending on age. Participants are given credit for each adjacent pair of pictures they correctly place (i.e., if pictures in locations 7 and 8 are placed in that order and adjacent to each other anywhere, such as slots 1 and 2, one point is awarded), up to the maximum value for the sequence, which is one less than the sequence length. (That is, if 18 pictures are in the sequence, the maximum score on that trial is 17 – the number of adjacent pairs of pictures that exist). The test takes approximately seven minutes to administer. This test is recommended for ages 3-85.

Scoring Process: The PSMT is scored using IRT methodology. The number of adjacent pairs placed correctly for each of trials 1 and 2 is converted to a theta score, which provides a representation of the given participant's estimated ability in this episodic memory task. All normative standard scores are provided.

Interpretation: The PSMT is a measure of episodic memory, which involves the acquisition, storage and effortful recall of new information. It is considered a strong "fluid ability" measure, with performance reaching a peak in early adulthood and declining across the life span. Measures of episodic memory such as PSMT can be extremely useful in evaluating performance of those with potential neurological impairments or other health-related problems in which memory is implicated or at risk. One can evaluate all three types of standard scores to interpret individual performance, with higher scores representing better episodic memory within the normative standard being applied (i.e., in relation to the general child or adult population, or in relation to age peers, or in relation to overall

demographically comparable peers). In addition to the three standard scores provided, the Uncorrected Standard score (or the theta score, alternatively) also provides a gauge of improvement or decline from one assessment to another, irrespective of demographic factors. A change in a participant's Uncorrected Standard score (or theta score) from Time 1 to Time 2 represents an absolute change in the level of performance for that individual since the previous assessment.

NIH Toolbox List Sorting Working Memory Test (List Sorting)

Description: The List Sorting test requires immediate recall and sequencing of different visually and orally presented stimuli (i.e., "working memory"). Pictures of different foods and animals are displayed with accompanying audio recording and written text (e.g., "elephant"), and the participant is asked to say the items back in size order from smallest to largest, first within a single dimension (either animals or foods, called 1-List) and then on two dimensions (foods, then animals, called 2-List). The test takes approximately seven minutes to administer and is recommended for ages 7-85, though a supplemental test designed for ages 3-6 is also available, if desired. This supplemental test also provides normative scores, though it may not appropriate for all children in this age range (especially 3-year-olds).

Scoring Process: List Sorting is scored by summing the total number of items correctly recalled and sequenced on 1-List and 2-List, which can range from 0-26. This score is then converted to the nationally normed standard scores described above.

Interpretation: List Sorting is a measure of working memory, tapping both information storage and processing (manipulation). It is considered a "fluid ability" measure, with performance tending to peak in early adulthood and then declining across the life span. The raw score obtained is converted to Age-Corrected and Uncorrected Standard Scores and Fully Corrected T-Scores based on the NIH Toolbox nationally representative sample, and associated percentiles can be looked up in this manual's Appendix A. Higher scores on each of these indicate higher levels of working memory within the normative standard being applied. To evaluate simple improvement or decline over time, one can also use the List Sorting raw score obtained on each assessment. It is important to note that the raw change in score from one assessment to another may not be consistent with a norm-referenced comparison of the individual's performance relative to peer groups.

NIH Toolbox Pattern Comparison Processing Speed Test (Pattern Comparison)

Description: This test measures speed of processing by asking participants to discern, as quickly as possible, whether two side-by-side pictures are the same or not. The items are presented one pair at a time on the iPad screen, and the participant is given 85 seconds of response time (excluding any time needed for the given iPad to "load" the items) to respond to as many items as possible (up to a maximum of 130). The items are designed to be simple so as to most purely measure processing speed. Overall, the test takes approximately three minutes to administer. This test is recommended for ages 7-85.

Scoring Process: The participant's raw score is the number of items answered correctly in 85 seconds of response time, with a range of 0-130. This score is then converted to the NIH Toolbox normative standard scores.

Interpretation: The Pattern Comparison Test is a measure of speed of processing, which typically improves steadily (time to complete task decreases) throughout childhood and adolescence, then begins to decline in adulthood, becoming much slower in older adults. As such, it is considered a "fluid ability" measure. The raw score obtained is converted to Age-Corrected and Uncorrected Standard Scores, Fully

Corrected T-Scores and associated Percentiles (see Appendix A) based on the NIH Toolbox normative sample. Higher scores indicate faster speed of processing within the normative standard being applied. To evaluate simple improvement or decline over time, one can use the raw score (range = 0-130) obtained on each assessment.

Slowed processing speed has been associated with normal aging, with decreases in processing speed being a significant contributor to age-related decline in other cognitive domains. Processing speed declines have also been found to impact several aspects of mental functioning in older age groups, including driving skills. Processing speed has also been shown to be highly vulnerable to brain damage, and multiple clinical populations demonstrate diminished processing speed. Assessments of processing speed have been found consistently to be among the most sensitive of neuropsychological measures (along with measures of episodic memory); typically, measures of processing speed are able to differentiate between clinical groups and healthy groups.

Cognition Supplemental Measures

NIH Toolbox Auditory Verbal Learning Test (Rey)

Description: The Rey is a word-list learning task in which 15 unrelated words are presented orally (via audio recording) over three consecutive learning trials[1]. After each presentation, the participant is asked to recall as many of the words as he/she can. The Rey is one of the most widely studied measures of memory and has been used in different languages, cultures and ethnic groups around the world. The test is typically administered to ages 8-85 (though it can be administered at other ages, at the researcher/clinician's discretion) and takes approximately three minutes. It can be administered as a supplement to the PSMT for even more detailed study of episodic memory, or as an accommodation in place of PSMT for those with significant visual impairment.

Scoring Process: The Rey is scored by taking the sum of the number of words recalled across all trials (possible range is 0-45 words). Toolbox norms and scale scores are *not available* for the Rey; however, descriptive statistics obtained from the sample of participants administered the Rey during the Toolbox norming study are available in the *NIH Toolbox Technical Manual*.

Interpretation: The raw score is most commonly used for interpretation of the Rey test, with higher scores representing better episodic memory. The test is sensitive to change over time and, like PSMT, can be useful in evaluating memory impairments or changes associated with many neurological disorders. One can also use the descriptive statistics provided in the *NIH Toolbox Technical Manual* to compare an individual's performance with others from the Toolbox norming sample; specific instructions are provided in the technical manual for such analyses.

[1]Many other versions of this test are available, but this description applies to the NIH Toolbox version. The NIH Toolbox version of this test differs from some other available versions in that three rather than five learning trials are provided, and neither a delayed recall trial nor an interference trial is given.

NIH Toolbox Oral Symbol Digit Test

Description: In this test, a coding key with nine abstract symbols is presented – each paired with a number between 1 and 9. Participants are asked to orally indicate which numbers go with symbols that are presented on a sheet of paper. The participant is given 120 seconds to call out as many numbers that go with the corresponding symbols as he/she can – without skipping any. This test is typically administered to ages 8-85 (though it can be administered at other ages, at the researcher/clinician's discretion) and takes approximately three minutes. The Oral Symbol Digit Test is a measure of processing speed. It can be administered as an accommodation in place of the Pattern Comparison Processing Speed Test for those with significant motor limitations in the upper extremities.

Scoring Process: The Oral Symbol Digit Test is scored as the number of items answered correctly in 120 seconds (possible range is 0-144). Toolbox norms and scale scores are *not available* for this test; however, descriptive statistics obtained from the sample of participants administered the test during the Toolbox norming study are available in the *NIH Toolbox Technical Manual*.

Interpretation: The raw score is most commonly used for interpretation of the Oral Symbol Digit Test, with higher scores representing better processing speed. One can also use the descriptive statistics provided in the *NIH Toolbox Technical Manual* to compare an individual's performance with others from the Toolbox norming sample; specific instructions are provided in the technical manual for such analyses.

Cognition Batteries and Composite Scores

The NIH Toolbox Cognition Battery for ages 7-85 includes all seven core measures described above. For ages 3-6, the NIH Toolbox Early Childhood Battery is recommended. This battery includes the Picture Vocabulary, Flanker, Dimensional Change Card Sort and Picture Sequence Memory measures, and takes approximately 20 minutes to administer to young children. In addition to scores for individual measures as described above, the cognition battery provides composite scores, which allow for general interpretation/evaluation of overall cognitive functioning and an even higher level of reliability than is possible with any individual test. To calculate a composite score, valid scores must be present for all component measures. The Early Childhood Battery produces one composite score, while the Cognition Battery for ages 7-85 produces three composite scores. The composite scores provided are as follows:

NIH Toolbox Fluid Cognition Composite Score

This composite includes all the tests noted above that are fluid ability measures: Flanker, Dimensional Change Card Sort, Picture Sequence Memory, List Sorting and Pattern Comparison. This composite score is derived by averaging the standard scores of each of the measures, and then deriving standard scores based on this new distribution. An Age-Corrected Standard Score, Fully Corrected T-Score, Uncorrected Standard Score and associated Percentiles (see Appendix A) are provided for the Fluid Cognition Composite.

Interpretation: One can interpret the Fluid Cognition Composite as a more global assessment of individual and group fluid cognition functioning. Higher scores indicate higher levels of functioning. An uncorrected or age-corrected standard score at or near 100 indicates ability that is average compared with others nationally (the comparison vs. age peers or the general population will depend on which standard score is being discussed). Standard scores around 115 suggest above-average fluid cognitive ability, while scores around 130 suggest superior ability (in the top 2 percent nationally, based on NIH Toolbox normative data). Conversely, a standard score around 85 suggests significantly below-average

fluid cognitive ability, and a score in the range of 70 or below suggests very low functioning, which may also be indicative of difficulties in school (for children) or general functioning. A Fully Corrected T-Score at or near 50 indicates ability that is average compared with others nationally and with similar demographic characteristics, and one below 40 suggests the possibility of health-related, acquired cognitive impairment.

Fluid abilities are used to solve problems, think and act quickly, and encode new episodic memories; they play an important role in adapting to novel situations in everyday life. Fluid abilities are considered to be especially influenced by biological processes and are less dependent on past exposure (learning experiences). These abilities improve rapidly during childhood, typically reaching their peak in early adulthood, then decline as adults get older. Fluid abilities tend to be more sensitive to neurobiological integrity, including changes in brain functioning with aging and in a variety of neurological, or systemic, disorders that alter brain structure and function. As such, the Fully Corrected Fluid Cognition Composite Score is recommended for use in studies of such conditions.

NIH Toolbox Crystallized Cognition Composite Score

This composite includes the Picture Vocabulary and Reading Tests. The composite score is derived by averaging the standard scores of each of the measures, and then deriving standard scores based on this new distribution. An Age-Corrected Standard Score, Fully Corrected T-Score, Uncorrected Standard Score and associated Percentiles (see Appendix A) are available for the Crystallized Cognition Composite.

Interpretation: One can interpret the Crystallized Cognition Composite as a more global assessment of individual and group verbal cognition. Higher scores indicate higher levels of functioning. A standard score at or near 100 indicates ability that is average compared with others nationally (the comparison vs. age peers or the general adult population will depend on which standard score is being discussed). Standard scores around 115 suggest above-average crystallized cognitive ability, while scores around 130 suggest superior ability (in the top 2 percent, based on NIH Toolbox normative data). Conversely, a standard score around 85 suggests below-average crystallized cognitive ability, and a score in the range of 70 or below suggests very low language skills, which may also be indicative of difficulties in school (for children) or general functioning in work environments with more of a verbal load. A Fully Corrected T-Score at or near 50 indicates ability that is average compared with others nationally and with similar demographic characteristics.

Crystallized abilities are presumed to be more dependent on experience and less on biological influences. They represent an accumulated store of verbal knowledge and skills, and thus are more heavily influenced by education and cultural exposure, particularly during childhood. These abilities show marked developmental change during childhood; they typically continue to improve slightly into middle adulthood, and then remain relatively stable.

NIH Toolbox Cognitive Function Composite Score

The Cognitive Function Composite is based upon an average of the Fluid and Crystallized composites. This composite score is derived by averaging the Fluid and Crystallized standard scores, then deriving standard scores based on this new distribution. An Age-Corrected Standard Score, Fully Corrected T-Score, Uncorrected Standard Score and associated Percentiles (see Appendix A) are available for the Cognitive Function Composite.

Interpretation: The Cognitive Function Composite Score can be interpreted much like a “full-scale score” in any commercially available test. It provides a highly reliable overall snapshot of general cognitive

functioning that may be of particular interest to researchers whose focus may be in other areas, but who want a strong general measure of cognition for individuals and groups. Higher scores indicate higher levels of cognitive functioning. A standard score at or near 100 indicates ability that is average compared with others nationally (the comparison will depend on which standard score is being discussed). Standard scores around 115 suggest above-average cognitive ability, while scores around 130 suggest superior ability (in the top 2 percent nationally, based on Toolbox normative data). Conversely, a standard score around 85 suggests below-average cognitive ability, and a score in the range of 70 or below (bottom 2 percent) suggests very low cognitive functioning, which may also be indicative of difficulties in school (for children) or general functioning. A Fully Corrected T-Score at or near 50 indicates ability that is average compared with others nationally and with similar demographic characteristics.

NIH Toolbox Early Childhood Composite Score

The Early Childhood Composite Score is derived from the four cognition measures that comprise the Early Childhood Battery: Picture Vocabulary, Flanker, DCCS and Picture Sequence Memory. It is designed primarily for NIH Toolbox users assessing children ages 3-6. This composite score is derived by averaging the standard scores of each of the four component measures, and then deriving standard scores based on this new distribution. An Age-Corrected Standard Score, Fully Corrected T-Score, Uncorrected Standard Score and associated Percentiles (see Appendix A) are available for the Early Childhood Composite.

Interpretation: The Early Childhood Composite Score can be interpreted much like the Cognitive Function Composite is for older children and adults. It provides a highly reliable overall snapshot of general cognitive functioning. Higher scores indicate higher levels of cognitive functioning. A standard score at or near 100 indicates ability that is average compared with others nationally (the comparison will depend on which standard score is being discussed). Standard scores around 115 suggest above-average cognitive ability, while scores around 130 suggest superior ability (in the top 2 percent nationally, based on NIH Toolbox normative data). Conversely, a standard score around 85 suggests below-average cognitive ability, and a score in the range of 70 or below suggests very low cognitive abilities (below the 2nd percentile), which may also be predictive of difficulties in school. A Fully Corrected T-Score at or near 50 indicates ability that is average compared with others nationally and with similar demographic characteristics.

NIH Toolbox® Motor Domain

The Motor domain measures Motor function, the ability to physically perform tasks, which is integrally related to daily functioning and quality of life. Five subdomains critical for optimal functioning across the life span were identified:

- locomotion - an act of moving from one place to another
- balance - the ability to orient body parts in space and maintain an upright posture under both static and dynamic conditions
- dexterity - an individual's ability to coordinate the fingers and manipulate objects in a timely manner
- strength - the capacity of a muscle to produce the tension and power necessary for maintaining posture, initiating movement, or controlling movement
- endurance - the ability to sustain effort that requires conjoint work capacities from cardiopulmonary function, biomechanical and neuromuscular function

Specific recommended age bands for each measure are noted below. Five core NIH Toolbox motor measures are available. There are no Supplemental motor measures.

Motor Core Measures

NIH Toolbox 9-Hole Pegboard Dexterity Test

Description: This simple test of manual dexterity records the time required for the participant to accurately place and remove nine plastic pegs into a plastic pegboard. The protocol includes one practice and one timed trial with each hand. Raw scores are recorded as time in seconds it takes the participant to complete the task with each hand (separate score for each). The test takes approximately four minutes to administer and is recommended for ages 3-85.

Scoring Process: The 9-Hole Pegboard Dexterity Test provides a score for each hand, with the primary NIH Toolbox score being the number of seconds it takes the participant to complete the task using his/her dominant hand ("handedness" is assessed at the outset of NIH Toolbox testing). This score is then converted to the NIH Toolbox normative scores. The non-dominant hand score is also reported as a raw score, showing number of seconds for completion, and NIH Toolbox normative scores are also provided, for those researchers interested in this additional information. This non-dominant hand information is provided in separate, "non-dominant"-labeled columns of the Toolbox Assessment Scores Report output file.

Interpretation: Dexterity is a central component of hand function and relates to both the speed and accuracy of hand movements during the manipulation of objects. For the 9-Hole Pegboard Dexterity Test, the raw score is commonly used for interpretation, with faster completion times (less time to complete) representing better manual dexterity. This also allows for raw score comparisons between dominant and non-dominant hand performance. However, one can also evaluate performance with the dominant and/or non-dominant hand by looking at the normative standard scores provided. For NIH Toolbox Motor assessment, the Fully Corrected T-Score is the score that should be primarily utilized for the interpretation of normative scores because it takes into account gender, age, ethnicity and education differences. Thus, it provides a level playing field for evaluating participants' performance since differences may exist in performance as a function of some of these demographic variables (most notably, gender and age).

When interpreting dexterity normative standard scores, higher performance is indicative of better dexterity. A Fully Corrected T-Score that is 2 SDs below the mean (30 or below) is suggestive of motor dysfunction; further evaluation by a physician or physical therapist is recommended. From an age perspective, dexterity in children is correlated with school performance and is a predictor of quality of handwriting, while a decline in manual dexterity is a common phenomenon in older adults and is associated with performance of activities of daily living and independent living.

NIH Toolbox Grip Strength Test

Description: This protocol is adapted from the grip strength testing protocol of the American Society of Hand Therapy. Participants are seated in a chair with their feet touching the ground. With the elbow bent to 90 degrees and the arm near the trunk, wrist at neutral, participants squeeze the Jamar Plus Digital dynamometer as hard as they can for a count of three. The dynamometer provides a digital reading of force in pounds. A practice trial at less than full force and one test trial are completed with each hand. The test takes approximately three minutes to administer and is recommended for ages 3-85.

Scoring Process: The Grip Strength Test provides a score for each hand, with the primary NIH Toolbox score being the number of pounds of force the participant was able to generate using his/her *dominant hand* ("handedness" is assessed at the outset of NIH Toolbox testing). This score is then converted to the NIH Toolbox normative scores. The non-dominant hand score is also reported as a raw score, in pounds of force, and NIH Toolbox normative scores are also provided, for those researchers interested in this additional information. This non-dominant hand information is provided in the separate, "non-dominant"-labeled columns of the Toolbox Assessment Scores Report output file.

Interpretation: Muscle strength is an essential element for humans to move against gravity and provide sufficient force to perform movements within the full range of motion. For the Grip Strength Test, the raw score has commonly been used for interpretation, with greater force (in pounds) representing greater strength. This also allows for raw score comparisons between dominant and non-dominant hand performance. However, one can also evaluate performance with the dominant and/or non-dominant hand by looking at the normative standard scores provided. For NIH Toolbox Motor assessment, the Fully Corrected T-Score is the score that should be primarily utilized for the interpretation of normative scores because it takes into account gender, age, ethnicity and education differences. Thus, it provides a level playing field for evaluating participants' performance since differences in performance may exist as a function of some of these demographic variables (most notably, gender and age). When interpreting strength normative standard scores, higher performance is indicative of better strength. A Fully Corrected T-Score that is 2 SDs below the mean (score of 30 or below) is suggestive of motor dysfunction; further evaluation by a physician or physical therapist is recommended. More generally, grip strength has been used to characterize total body strength and predict mortality, postsurgical complications and future disability. Muscle strength of the limbs and trunk declines with age and is associated with an increased risk of falls, hip fractures, loss of bone mineral density, long-term survival in severe congestive heart failure, functional dependence in people aged 75 years or older, and loss of functional status in hospitalized patients.

NIH Toolbox Standing Balance Test

Description: The Standing Balance Test is a measure developed to assess static standing balance for ages 3-85 years. It involves the participant assuming and maintaining up to five poses for 50 seconds each. The sequence of poses is: eyes open on a solid surface, eyes closed on a solid surface, eyes open on a

foam surface, eyes closed on a foam surface, eyes open in tandem stance on a solid surface. Detailed stopping rules are in place to ensure participant safety with these progressively demanding poses. Postural sway is recorded for each pose using an iPod Touch that the participant wears at waist level. This test takes approximately seven minutes to administer and is recommended for ages 3-85.

Scoring Process: The participant's anterior-posterior postural sway information is fed wirelessly to the iPad. A normalized path length score is then calculated as follows:

$$Normalized\ Path\ Length = \frac{1}{t}\sum_{j=1}^{N-1}|p_{j+1} - p_j|$$

Where *t* is the time duration, *N* is the number of time samples, and p_j is accelerometer data at time sample *j*. These data are then further converted using an IRT model to derive a theta score for each participant representing the relative overall balance ability or performance of the participant. Age-Corrected, Fully Corrected and Uncorrected Standard Scores are then provided for the Standing Balance Test, and associated percentiles (see Appendix A) are available. In addition, two ratio scores are provided, comparing performance on balance position 2 to position 1, and position 4 to position 1. These ratios can provide some potentially useful information for clinicians in evaluating certain subjects' risk of falling. These ratios are labeled columns on the NIH Toolbox Assessment Scores Report output file.

Interpretation: Balance allows humans to be able to orient the body in space, maintain an upright posture under static and dynamic conditions, and move without falling. To evaluate motoric balance with the Standing Balance Test, one can look at the normative standard scores provided. For NIH Toolbox Motor assessment, the Fully Corrected T-Score is the score that should primarily be utilized for the interpretation of normative scores because it takes into account gender, age, ethnicity and education differences. Thus, it provides a level playing field for evaluating participants' performance since differences in performance may exist as a function of some of these demographic variables (most notably, gender and age). When interpreting balance normative standard scores, higher performance is indicative of better balance. A Fully Corrected T-Score that is 2 SDs below the mean (score of 30 or below) is suggestive of motor dysfunction; further evaluation by a physician or physical therapist is recommended. When evaluating ratio scores, the position 2/position 1 ratio represents the participant's ability to use input from the somatosensory and vestibular systems to maintain balance, while the position 4/position 1 ratio reflects the relative reduction in postural stability when visual and somatosensory inputs are simultaneously disrupted (typically representative of the effectiveness of vestibular function for postural control). Generally, lower ratio scores (those closer to 1) are better. As noted, these may be of use to clinicians. Examination of balance is important as it predicts a person's ability to safely and independently function in a variety of environments. Maintaining stance stability under varying sensory environments is an essential function for the elderly to avoid falling and among patients for better functional outcomes. Several studies have found that changes in balance ability correlate significantly with changes in function.

NIH Toolbox 4-Meter Walk Gait Speed Test

Description: This test is adapted from the 4-meter walk test in the Short Physical Performance Battery. Participants are asked to walk a short distance (four meters) at their usual pace. Participants complete one practice and then two timed trials. Raw scores are recorded as the time in seconds required to walk 4 meters on each of the two trials, with the better trial used for scoring. The test takes approximately

three minutes to administer (including instructions and practice). This test is recommended for ages 7-85.

Scoring Process: The raw score on the 4-Meter Walk Test is the number of seconds it takes to walk four meters, using the better of two trials. This is then transformed into a computed score reported in meters per second. For example, if it took a participant two seconds to walk four meters, one would divide four by two to get two meters per second as the score. NIH Toolbox normative scores are *not available* for this test; however, descriptive statistics obtained from the sample of participants who were administered a version of the test during the NIH Toolbox norming study are available in the *NIH Toolbox Technical Manual*.

Interpretation: On the 4-Meter Walk Gait Speed Test, higher computed scores are indicative of better gait speed (i.e., fewer seconds to walk four meters). One can evaluate the descriptive statistics in the *NIH Toolbox Technical Manual* to get a sense of how individual or group performance compares to results obtained from the national norming sample, though care should be exercised in specific interpretation. Gait speed as a measure of bipedal locomotion is both a good way to summarize the overall burden of disease as well as a generic indicator of health status, prognosis and the co-morbid burden of disease in older persons. The speed at which older individuals walk is relevant to their functioning in the community. Moreover, gait speed is an important predictor of outcomes such as: length of stay and discharge disposition of patients admitted for acute rehabilitation after stroke, mortality, incident ischemic stroke and incident dementia.

NIH Toolbox 2-Minute Walk Endurance Test

Description: This test is adapted from the American Thoracic Society's 6-Minute Walk Test Protocol. This test measures sub-maximal cardiovascular endurance by recording the distance that the participant is able to walk on a 50-foot (out and back) course in two minutes. The participant's raw score is the distance in feet and inches walked in two minutes. The test takes approximately four minutes to administer (including instructions and practice). This test is recommended for ages 3-85.

Scoring Process: The participant's raw score is the distance walked in two minutes, reported in feet (and fractions thereof). This score is then converted to the NIH Toolbox normative scores.

Interpretation: Cardiorespiratory and muscle endurance are important components of physical fitness and contribute to both performance and health status. On the 2-Minute Walk Endurance Test, greater distance walked is suggestive of better endurance. To evaluate endurance with this test, one can look at the normative standard scores provided. For NIH Toolbox Motor assessment, the Fully Corrected T-Score is the score that should be primarily utilized for the interpretation of normative scores, because it takes into account gender, age, ethnicity and education differences. Thus, it provides a level playing field for evaluating participants' performance since differences in performance may exist as a function of some of these demographic variables (most notably, gender and age). When interpreting endurance normative standard scores, higher performance is indicative of better endurance. A Fully Corrected T-Score that is 2 SDs below the mean (score of 30 or below) is suggestive of motor dysfunction; further evaluation by a physician or physical therapist is recommended. People with better endurance are able to complete daily tasks and are more fit to pursue leisure activities and accomplish higher-intensity workloads. The clinical significance of endurance as measured by timed walk tests to morbidity and mortality outcomes has been extensively reported in healthy and clinical populations across the age span.

Motor Batteries

The NIH Toolbox Motor Battery for ages 7-85 includes all five core measures described above. For ages 3-6, the NIH Toolbox Early Childhood Motor Battery includes four core tests, but excludes the 4-Meter Walk Gait Speed Test. Individual scores are provided for each measure, as described above, but no composite scores are provided for the Motor Battery.

NIH Toolbox® Sensation Domain

The Sensation Domain measures several aspects of sensory functioning for ages 3-85, referred to here as subdomains, including audition, vision, vestibular, olfaction and taste. In addition, two survey measures of pain are provided. Specific recommended age bands for each measure are noted below; available sensation batteries by age are described at the end of this section. For this section, each subdomain is described separately.

Sensation Subdomains and Measures

Audition

Audition (hearing) is the processing of sound in the environment. It is necessary for navigating in the environment and communicating with others. Acoustic information is processed through three groups of peripheral structures (outer, middle and inner ears) and then through the central auditory nervous system to create auditory experience. One core NIH Toolbox audition measure exists. Norms are not available for this measure, but useful scores and interpretive information are provided below. In addition, descriptive statistics obtained from the sample of participants who were administered the test during the NIH Toolbox norming study are available in the *NIH Toolbox Technical Manual*.

NIH Toolbox Words-in-Noise Test (WIN)

Description: This test measures a person's ability to recognize single words presented amid varying levels of background noise. It measures how much difficulty a person might have hearing in a noisy environment. A recorded voice instructs the participant to listen to and then repeat words. The task becomes increasingly difficult as the background noise gets louder, thus reducing the signal-to-noise ratio. The test is recommended for participants ages 6-85 and takes approximately six minutes to administer.

Scoring Process: The examiner scores the participant's responses as correct or incorrect, and a total raw score (out of a maximum of 35 points) is calculated by the software for each ear. A percent correct is calculated, which is then translated into a threshold score for each ear, in decibels of signal-to-noise ratio (dB S/N), using a look-up table (see Appendix B). Alternatively, the following equation can be used to calculate the S/N score based on the raw score, in lieu of the look-up table. For each ear:

```
WIN_Score = 26-0.8*WIN_NCorrect
```

Thus, the best score that can be attained (35 correct) for either ear is -2.0 dB S/N, and the worst score (0 correct) is 26.0 dB S/N. Lower scores, therefore, are indicative of better performance on this test. In the Toolbox Assessment Scores Report output file, threshold scores and raw scores are provided for each ear.

Interpretation: Assessment of the ability to understand speech in a noisy background yields an ecologically valid measure of hearing because a substantial portion of communication in the real world occurs in less-than-ideal environments. Moreover, speech perception in noise is often difficult to predict from pure-tone thresholds or from speech perception in quiet settings. The interpretive guidelines provided are preliminary and may need further adjustment as future studies are conducted.

As noted above, the range of possible scores for each ear is -2.0 to 26.0 dB S/N, with lower scores indicative of better performance and, conversely, higher scores potentially suggestive of hearing

difficulties. For score interpretation with ages 13 and above, a cutoff of 10 dB S/N is recommended for the NIH Toolbox version of this measure. Participants with a score higher than this cutoff should follow up with a hearing professional, specifically an otolaryngologist, who would then refer to an audiologist as needed. Users should note that the cutoff suggested here is slightly higher than other published versions of this test because other versions were conducted in quieter environments.

For score interpretation with children ages 6-12, different cutoffs are recommended, as follows:

Age	6	7	8	9	10	11	12
Suggested Cutoff (dB S/N)	16.2	13.0	13.0	11.4	11.4	11.4	11.4

Taste

Taste perception, also known as gustation, arises from stimulation of taste receptors composed of epithelial cells found most frequently on the papillae of the tongue throughout the oral cavity. One core NIH Toolbox taste measure is available; national normative scores are provided.

NIH Toolbox Regional Taste Intensity Test

Description: This test measures the perceived intensity of quinine (a bitter tastant) and salt administered in liquid solutions. The tastants are each applied to the tip of the tongue as well as swished around in the whole mouth and are rated on a generalized labeled magnitude scale (gLMS). The gLMS is a measure of perceived intensity, with seven anchor labels provided (*Strongest imaginable, Very strong, Strong, Moderate, Weak, Barely detectable, No sensation*). Participants can rate their intensity by touching the iPad screen at any point on the scale from Strongest imaginable to No sensation. The software records the exact location of the response. The test is recommended for administration to participants ages 12-85 and takes approximately six minutes to administer.

Scoring Process: A score from 0-100 on a semi-logarithmic scale is produced for each of the four items (quinine whole mouth, salt whole mouth, quinine tip of tongue, salt tip of tongue), corresponding to the point on the gLMS where the participant clicked. A higher score represents greater perceived intensity of the tastant. Normative scores and gLMS scores are provided for quinine whole mouth and salt whole mouth items in the NIH Toolbox Assessment Scores Report output file.

Interpretation: To evaluate perceived taste intensity with this test, one can look at the normative standard scores provided. When interpreting normative standard scores for whole mouth quinine or whole mouth salt items, higher performance is indicative of higher perceived taste intensity of the items. For research purposes, results can be grouped according to the level of intensity of taste perceived, which can then be evaluated in terms of other outcome variables. In addition, from a clinical perspective, participants who score at or below the 10th percentile (very low perceived taste intensity) on either the Age-Corrected or Fully Corrected T-Scores for whole mouth quinine or salt might be flagged for referral to a physician.

Vision

Vision is a complex sensation that provides us with a personal conscious representation of our surrounding environment. Loss of vision or blindness may limit a person's ability to complete normal,

daily activities and decrease overall quality of life. Visual impairment can impose various limitations on a person's functional ability, including reading, mobility (which includes driving), visual information processing (also called "seeing"), and visually guided motor behavior (also called "manipulation"). One core NIH Toolbox vision measure is available with national normative scores provided.

NIH Toolbox Visual Acuity Test

Description: This test directly measures participants' visual acuity or distance vision. The participant is seated 3 meters away from the iPad screen at eye level, and letters (called "optotypes") are displayed one at a time on the screen for the participant to identify, using both eyes at the same time, with the participant wearing his/her normal corrective lenses for distance vision (glasses or contact lenses), if worn. As the participant successfully identifies optotypes of a given size, smaller ones appear on the screen, until the software ascertains the smallest-size optotype the participant can successfully see. Conversely, the program displays larger optotypes if the participant cannot see the size that is first displayed, until a size that he/she can accurately see is found. For participants ages 3-7, only the letters H, O, T and V are used, and children may point to a laminated card showing the letters if they cannot verbalize or recall the letter names. For participants ages 8 and above, the entire set of optotypes is used, following a common protocol used in professional vision testing. This test takes approximately three minutes to administer and is recommended for ages 3-85.

Scoring Process: This is the standard binocular visual acuity measure scored in LogMAR units. The reciprocal of the Snellen notation (most often cited by lay individuals) equals the angle (in minutes of arc), which the strokes of the letter subtend at the subject's eye. It is called the minimum angle of resolution (MAR). LogMAR is MAR expressed in $\log_{10}$ form. For example, a Snellen VA of 20/200 has a MAR of 10 and a LogMAR of 1; a Snellen VA of 20/20 has a MAR of 1 and a LogMAR of 0. The LogMAR value can be calculated from the raw number correct (ranging from 0-95) according to the following equation:

$$1.6 - (0.02*RAW).$$

Therefore, the worst possible LogMAR score is 1.6, while the best score possible on this test is -0.3. As noted above, the software adjusts the size of the optotype presented to most efficiently measure the participant's true visual acuity, so not all items need to be administered; credit is given for larger sizes not administered.

The LogMAR score on this test is provided in the Computed Score field of the Assessment Scores Report output file. For NIH Toolbox users who wish to know the Snellen equivalent for a given participant's performance, a value is provided in the Assessment Scores Report output file, labeled Static Visual Acuity Snellen. For more detailed information, an equivalency table is provided in Appendix C. Snellen values range from 20/10 (highest acuity measured) to 20/640 (lowest acuity measured by NIH Toolbox iPad app). In addition to LogMAR scores, NIH Toolbox normative scores are provided for the Visual Acuity Test.

Interpretation: The Visual Acuity Test provides a reliable measure of participants' overall functional distance vision since it measures both eyes simultaneously. Users can monitor participants' change over time in LogMAR units or can evaluate participants' normative performance using NIH Toolbox standard scores. Normative scores can be useful if one wishes to evaluate an individual's or group's relative vision performance – that is, does this person have average (or below or above average) distance vision compared to the national population, based on age or many demographic factors? This is just one example of how vision normative scores could be utilized. In everyday life, a Snellen equivalent of 20/40

or better (including corrective lenses) is typically a requirement of obtaining a motor vehicle license. Federal government definitions of “legally blind” refer to corrected vision in the individual’s better eye of 20/200 and worse. For any NIH Toolbox Visual Acuity Test score worse than a Snellen equivalent of 20/40 with best correction, a referral to an eye care professional is recommended.

Olfaction

The primary purpose of the olfactory system in humans is to detect and perceive volatile airborne chemicals and thus provide information about our environment and food quality that is critical to our health, a nutritious diet and psychological well-being. One core NIH Toolbox olfaction measure is available; national normative scores are provided.

NIH Toolbox Odor Identification Test

Description: This task assesses a person’s ability to identify various odors. Participants use scratch-and-sniff cards and after scratching them one at a time, are asked to identify which of four pictures on the iPad screen matches the odor they have just smelled. Participants ages 10-85 are administered nine odor cards, while those ages 3-9 are administered five odor cards. Child participants (ages 3-9 years) are first asked to identify the eight pictures used as answer choices to ensure they can complete the task. Having identified the pictures, they are asked if they have tasted or smelled the objects or foods depicted. This test takes approximately four to five minutes to administer and is recommended for ages 3-85.

Scoring Process: Scores are calculated by simply summing the total number of correct items (score range for ages 3-9 is 0 to 5; for ages 10+, it is 0 to 9). Normative scores are provided based on this raw score.

Interpretation: Olfactory testing can be a useful adjunct to comprehensive assessments of health and well-being. For example, impaired olfactory function is now recognized as one of the hallmark early signs of several neurodegenerative disorders, including Alzheimer’s and Parkinson’s disease. Nevertheless, individuals can vary widely in their ability to detect, recognize and identify odors, yet still be within the range of normal function. When evaluating normative standard scores for the Odor Identification Test, higher scores indicate better olfactory ability/functioning. If one uses the Age-Corrected or Fully Corrected T-Scores to look more closely at comparative performance based on age and other demographic factors, a concern about performance might be raised for participants scoring more than 1 SD below the mean (standard score below 85; T-score below 40). Such individuals may warrant further evaluation or follow-up, depending on whether any concurrent sinus or other related conditions were reported concurrently by the participant, as these could interfere with the ability to identify odors. If one were to evaluate raw score performance, a decline in performance would be expected from early adulthood through old age. Additional work remains to examine the predictive validity of this measure and to identify clinically meaningful thresholds.

Pain

Pain is an important component of health and function. Pain has been defined as an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage. Pain is a major symptom in many medical conditions and can significantly interfere with a person's quality of life and general functioning. The NIH Toolbox measures of pain focus on a participant’s reported pain experience, as well as the intensity of the pain experienced. Both measures used are derived from the published NIH Patient Reported Outcome Measurement Information System

(PROMIS)® instruments. NIH Toolbox norms and standard scores are *not available* for the pain measures; however, descriptive statistics obtained from the sample of participants who were administered the pain scales during the NIH Toolbox norming study are available in the *NIH Toolbox Technical Manual*. In addition, the NIH Toolbox Pain Interference Survey offers a T-Score based on the PROMIS sample (see below).

NIH Toolbox Pain Intensity Survey

Description: This measure consists of a single item measuring immediate (i.e., acute) pain in adults. It asks a participant to rate level of pain experienced "over the last seven days." It takes less than one minute to administer and is recommended for ages 18-85.

Scoring Process: The single item is simply scored on a 0-10 scale, with 0 representing no pain, and 10 representing the "worst imaginable pain." No derived scores are available.

Interpretation: One could reasonably expect a large proportion of the normal population to obtain scores of zero on this measure. Regardless, it is an easily quantifiable piece of information on one's subjective pain experience. Additional work remains to examine the predictive validity of this measure and to identify any clinically meaningful thresholds.

NIH Toolbox Pain Interference Survey

Description: This brief self-report scale measures the degree to which pain interferes with other activities in life in adults. Pain interference items were developed as part of the NIH PROMIS. This measure is administered as a CAT and takes approximately three minutes. It is recommended for ages 18-85.

Scoring Process: Each item administered has a 5-point scale with options ranging from "not at all" to "very much" on questions about how much pain interferes with aspects of one's life. The survey is scored using IRT methods. An IRT theta score is generated for each participant, and while no NIH Toolbox norms are available for this measure, the IRT scores are converted to general T-scores based on the PROMIS sample to which this test was given. These scores are provided in a column labeled "T-Score" on the Assessment Scores Report output file.

Interpretation: The pain interference item bank measures the self-reported consequences of pain on relevant aspects of one's life. This includes the extent to which pain hinders engagement with social, cognitive, emotional, physical and recreational activities. Higher theta and T-Scores represent greater participant report of pain interference in daily life. Thus, T-Scores ≥ 60 may be of concern. Additional work remains to examine the predictive validity of this measure and to identify any clinically meaningful thresholds.

Sensation and Pain Batteries

There are no NIH Toolbox Sensation and Pain batteries currently published for the iPad. Future releases may include such batteries.

NIH Toolbox® Emotion Domain

Emotion refers to any strong feelings, such as joy, sorrow or fear. Emotion is an affective state of consciousness in which joy, sorrow, fear, hate or the like is experienced, as distinguished from cognitive and volitional states of consciousness. Emotions can be negative and distressing, or positive emotions can be reflections of well-being in our lives. Positive social relationships can buffer stress and enhance health. Recognizing the full spectrum of emotional life and its impact on health, the mandate for the NIH Toolbox was to develop assessments with a broad focus, beyond just negative emotion or emotional distress. It includes additional aspects of the experience and expression of emotion relevant to general health, including the importance of psychological well-being, the role of important aspects of positive functioning such as adaptability, resilience and self-efficacy, and the importance of the interpersonal and social context in which emotions arise and may be expressed. Emotional health has significant links to physical health and exerts a powerful effect upon perceptions of life quality.

Four central subdomains are assessed in the Emotion domain: Psychological Well-Being, Social Relationships, Stress and Self-Efficacy and Negative Affect. Within each of these subdomains, specific concept areas are measured and, in some cases, sub-concepts have their own specific measures. Different versions of the Emotion surveys are available for different ages, as well as parent-report versions for young children because they are not able to respond directly to written surveys in most cases. Specific recommended age bands for each measure are noted below; available Emotion batteries by age are described at the end of this section. For this section, each subdomain is described separately.

For each measure, an Uncorrected Standardized Score (T-Score) is provided for the respondent, utilizing a normative mean of 50, with an SD of 10. For ages 3-17 (including parent report measures), this score uses a mean of 100 and an SD of 15 (scores can be compared to T-scores using Appendix A). In addition, for measures administered at ages 3-17 (whether self- or parent-report), an Age- and Gender-Corrected T-Score is provided. As the name indicated, this latter T-Score for Emotion measures corrects for age and gender only. Associated Percentiles may also be found for each T-Score in Appendix A.

Emotion Subdomains and Measures

Psychological Well-Being

Psychological well-being includes both hedonic and eudaimonic aspects of well-being. Hedonic aspects are more subjective and experiential and emphasize pleasure and positive affect (happiness, serenity and cognitive engagement). Eudaimonic well-being is more evaluative in nature and emphasizes fulfillment and purpose (e.g., meaning, life satisfaction). NIH Toolbox includes measures for three components of psychological well-being.

Positive Affect refers to feelings that reflect a level of pleasurable engagement with the environment, such as happiness, joy, excitement, enthusiasm and contentment. It is measured by the NIH Toolbox Positive Affect Survey.

NIH Toolbox Positive Affect Survey

Description: This self-report measure assesses both activated (i.e., happiness, joy) as well as unactivated (i.e., serenity, peace) aspects of positive affect. CAT versions are used for ages 13-17 and 18-85; a 9-item fixed-length form is used for ages 8-12, and a CAT is used for the parent-report versions for ages 3-7 and 8-12.

Scoring Process: Each item administered has a 5-point scale with options ranging from “not at all” to “very much.” Each survey is scored using IRT methods. An IRT-derived theta score is generated for each

participant, as well as an Uncorrected Standard Score (T-Score). In addition, for ages 3-17, an Age- and Gender-Corrected T-Score is provided.

Interpretation: For the NIH Toolbox Positive Affect Survey, higher scores are indicative of more positive affect. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of positive affect and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of positive affect. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Life Satisfaction is one's cognitive evaluation of life experiences; this measure is concerned with whether people like their lives or not. Life satisfaction includes both general (e.g., my life is going well) and domain-specific (e.g., I am satisfied with my family life) aspects. The general aspect is measured by the NIH Toolbox General Life Satisfaction Survey, which is in the NIH Toolbox Emotion Battery.

NIH Toolbox General Life Satisfaction Survey

Description: This self-report measure assesses global feelings and attitudes about one's life. A CAT is used for adults, a CAT version is used for ages 13-17, and a 5-item fixed-length form is used for ages 8-12, as well as for the parent-report version with ages 3-12.

Scoring Process: Items administered include those with both 5-point and 7-point scales, with options in each case ranging from "strongly disagree" to "strongly agree." The self-report surveys are scored using IRT methods, whereas the parent-report version is scored as a raw sum. The IRT-derived theta score and the raw summed score are each converted to Uncorrected T- Scores (respectively). For ages 3-17, an Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox General Life Satisfaction Survey, higher scores are indicative of more general life satisfaction. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of general life satisfaction and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of general life satisfaction. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Meaning and Purpose is characterized by the extent to which people feel their life matters or makes sense. It is measured by the NIH Toolbox Meaning and Purpose Survey.

NIH Toolbox Meaning and Purpose Survey

Description: This is a self-report measure administered only to ages 18-85 as a CAT.

Scoring Process: Items administered use a 5-point scale, with options ranging from "strongly disagree" to "strongly agree," or from "not at all" to "very much." The survey is scored using IRT methods. The IRT-derived theta score is converted to an Uncorrected Standard Score (T-Score).

Interpretation: Higher scores indicate more self-reported meaning and purpose. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of meaning and purpose and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of meaning and purpose. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Social Relationships

Social relationships have several dimensions, including their structure, extent and quality. NIH Toolbox focuses on four aspects of social relationships.

Perceived social support is the extent to which an individual views his/her social relationships as available to provide aid in times of need or when problems arise. It includes instrumental and emotional types of perceived social support. Emotional Support refers to the perception that people in one's social network are available to listen to one's problems with empathy, caring and understanding. It is measured by the NIH Toolbox Emotional Support Survey. Instrumental Support refers to the perception that people in one's social network are available to provide material or functional aid in completing daily tasks, if needed; it is measured by the NIH Toolbox Instrumental Support Survey.

NIH Toolbox Emotional Support Survey

Description: This self-report measure assesses emotional support through two fixed-length forms: an 8-item form for ages 18-85 and a 7-item form for ages 8-17. No parent-report versions are available.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." Each survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, as is an Uncorrected Standard Score (T-Score). For ages 8-17, an Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Emotional Support Survey, higher scores are indicative of more emotional support. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of support, and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of support. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Instrumental Support Survey

Description: This self-report measure assesses instrumental support for ages 18-85, using an 8-item fixed-length form. No versions for other ages are available.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, as is an Uncorrected Standard Score (T-Score).

Interpretation: For the NIH Toolbox Instrumental Support Survey, higher scores are indicative of more reported support. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of instrumental support and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of instrumental support. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Companionship is characterized by self-reported perceptions of the availability of friends or companions with whom to interact or affiliate (i.e., friendship) and perceptions that one is alone, lonely or socially isolated from others (i.e., loneliness). Companionship is measured by the NIH Toolbox Friendship Survey, NIH Toolbox Loneliness Survey, NIH Toolbox Social Withdrawal Survey and NIH Toolbox Positive Peer Interaction Survey.

NIH Toolbox Friendship Survey

Description: This self-report measure assesses perceptions of friendship, using an 8-item fixed-length form for ages 18-85 and a 5-item fixed-length form for ages 8-17.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, which is then converted to an Uncorrected Standard Score (T-Score). For ages 8-17, an Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Friendship Survey, higher scores are indicative of a greater perceived availability of companions with whom to interact or affiliate. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of friendship and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of friendship. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Loneliness Survey

Description: This self-report measure assesses perceptions of loneliness using a 5-item fixed-length form for ages 18-85 and a 7-item fixed-length form for ages 8-17.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, as is an Uncorrected Standard Score (T-Score). For ages 8-17, an Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Loneliness Survey, higher scores are indicative of more loneliness. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of loneliness and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of loneliness. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Positive Peer Interaction Survey

Description: As a conceptual analogue to the NIH Toolbox Friendship Survey, the NIH Toolbox Positive Peer Interactions Survey is a parent-report measure for children ages 3-12. It is a 4-item fixed-length survey that assesses how often a child plays with friends and gets along with peers.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." Items are scored and summed; the raw summed score is converted to an Uncorrected Standard Score (T-Score). An Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Positive Peer Interaction Survey, higher scores are indicative of more positive peer interactions. Scores 1 SD or more below the mean (T ≤ 40) suggest fewer positive relationships and scores 1 SD or more above the mean (T ≥ 60) suggest more positive relationships. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Social Withdrawal Survey

Description: As an analogue to the NIH Toolbox Loneliness Survey, the NIH Toolbox Social Withdrawal Survey is a parent-report measure for children ages 3-12. It is a 4-item fixed-length survey that assesses how often a child avoids or withdraws from social activities with peers.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." Items are scored and summed; the raw summed score is converted to an Uncorrected Standard Score (T-Score). An Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Social Withdrawal Survey, higher scores are indicative of higher levels of social withdrawal. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of social withdrawal and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of withdrawal. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Social distress is the extent to which an individual perceives his/her daily social interactions as negative or distressing. This can include aspects of perceived hostility (e.g., how often people argue with me, yell at me, or criticize me) and perceived insensitivity (e.g., how often people don't listen when I ask for help, or do not pay attention to me). Self-reported perceived hostility is measured by the NIH Toolbox Perceived Hostility Survey; perceived insensitivity is measured by the self-report NIH Toolbox Perceived Rejection Survey and the parent-report NIH Toolbox Peer Rejection Survey.

NIH Toolbox Perceived Hostility Survey

Description: This self-report measure assesses perceptions of hostility using an 8-item fixed-length form for ages 18-85 and a 5-item fixed-length form for ages 8-17.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, which is then converted to an Uncorrected Standard Score (T-Score). For ages 8-17, an Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Perceived Hostility Survey, higher scores are indicative of greater perceived hostility. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of perceived hostility and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of perceived hostility. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Perceived Rejection Survey

Description: This self-report measure assesses perceptions of rejection using an 8-item fixed-length form for ages 18-85 and a 5-item fixed-length form for ages 8-17.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, and an Uncorrected Standard Score (T-Score) is provided. For ages 8-17, an Age- and Gender-Corrected T-Score is also included.

Interpretation: For the NIH Toolbox Perceived Rejection Survey, higher scores are indicative of greater perceived rejection. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of perceived rejection and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of perceived rejection. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Peer Rejection Survey

Description: As an analogue to the NIH Toolbox Perceived Rejection Survey, the NIH Toolbox Peer Rejection Survey is a parent-report measure for children ages 3-12. It is a 9-item fixed-length form that assesses how often a child is left out, avoided or teased by peers.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." Items are scored and summed; the raw summed score is converted to an Uncorrected Standard Score (T-Score). An Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Peer Rejection Survey, higher scores are indicative of greater peer rejection. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of peer rejection and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of peer rejection. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Positive Social Development is characterized by parents' evaluation of their children's empathic behaviors. It is an indicator of a child's current emotional health and a predictor of positive and supportive social relationships in adolescence and adulthood. It is measured by the NIH Toolbox Empathic Behaviors Survey.

NIH Toolbox Empathic Behaviors Survey

Description: This parent-report measure for children ages 3-12 assesses parent perceptions of children's prosocial behaviors using a CAT.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, as is an Uncorrected Standard Score (T-Score). An Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Empathic Behaviors Survey, higher scores are indicative of more parent-reported child prosocial behaviors. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of parent-reported child empathic behaviors and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of parent-reported child empathic behaviors. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Stress and Self-Efficacy

Stress and Self-Efficacy focuses on individual perceptions about the nature of events and their relationship to the perceived coping resources of an individual. In general, psychological stress is said to occur when an individual perceives that environmental or internal demands that are personally meaningful exceed his/her adaptive capacity. NIH Toolbox assesses two areas related to stress and adaptive capacity.

Perceived Stress is defined by individual perceptions about the nature of events and their relationship to the values and coping resources of an individual. It is measured by the NIH Toolbox Perceived Stress Survey.

NIH Toolbox Perceived Stress Survey

Description: This is a self-report measure administered to ages 13-17 and 18-85, as well as a parent-report measure for ages 8-12. The self-report versions are administered as 10-item fixed-length forms. The parent-report version is administered as a CAT. This measure assesses how unpredictable, uncontrollable and overloaded respondents find their lives.

Scoring Process: Items administered use a 5-point scale, with options ranging from "never" to "very often." Each survey is scored using IRT methods. The IRT-derived theta score is converted to an Uncorrected Standard Score (T-Score). For ages 8-17, an Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Perceived Stress Survey, higher scores are indicative of more perceived stress. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of perceived stress and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of perceived stress. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Self-Efficacy can be described as a person's belief in his/her capacity to manage functioning and have control over meaningful events. It is measured by the NIH Toolbox Self-Efficacy Survey.

NIH Toolbox Self-Efficacy Survey

Description: This is a self-report measure administered to ages 8-12, 13-17, and 18-85, as well as a parent-report measure for ages 8-12. All versions are administered as CAT. It assesses respondents' sense of global self-efficacy.

Scoring Process: Items administered use a 5-point scale, with options ranging from "never" to "very often." Each survey is scored using IRT methods. The IRT-derived theta score is converted to an Uncorrected Standard Score (T-Score). For ages 8-17, an Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Self-Efficacy Survey, higher scores are indicative of more general self-efficacy. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of general self-efficacy and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of general self-efficacy. T-scores ≤ 40 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Negative Affect

Negative affect is a phrase used to describe unpleasant feelings or emotions that exist on a continuum ranging from common and normal feelings of sadness, fear and anger to more extreme feelings along the same continuum. Negative affect is understood as comprising important underlying dispositions (e.g., neuroticism, negative emotional style) and more transient negative feeling states. The focus of this subdomain is on three principal negative emotions: anger, fear and sadness.

Anger is characterized by attitudes of hostility and cynicism and is often associated with experiences of frustration impeding goal-directed behavior. For adult self-report, anger is comprised of three components: anger as an emotion, aggression as a behavioral component, and hostility as a set of cynical attitudes and mistrust of others and their motives. Anger is measured by the NIH Toolbox Anger-Affect Survey, NIH Toolbox Anger-Hostility Survey and the NIH Toolbox Anger-Physical Aggression Survey. For children, anger is measured by the NIH Toolbox Anger Survey.

NIH Toolbox Anger-Physical Aggression Survey

Description: This self-report measure assesses aggression as a behavioral component for ages 18-85 using a 5-item fixed-length form. No versions for other ages are available.

Scoring Process: Each item administered has a 7-point scale with options ranging from "extremely untrue of me" to "extremely true of me." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, as is an Uncorrected Standard Score (T-Score).

Interpretation: For the NIH Toolbox Anger-Physical Aggression Survey, higher scores are indicative of more reported physical aggression. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of physical aggression and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of physical aggression. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Anger-Hostility Survey

Description: This self-report measure assesses attitudes of hostility and cynicism for ages 18-85 using a 5-item fixed-length form. No versions for other ages are available.

Scoring Process: Each item administered has a 7-point scale with options ranging from "extremely untrue of me" to "extremely true of me." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, and an Uncorrected Standard Score (T-Score) is provided.

Interpretation: For the NIH Toolbox Anger-Hostility Survey, higher scores are indicative of more hostility. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of hostility and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of hostility. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Anger-Affect Survey

Description: This self-report measure assesses anger as an emotion for ages 18-85, using a CAT format. No versions for other ages are available.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, which is then converted to an Uncorrected Standard Score (T-Score).

Interpretation: For the NIH Toolbox Anger-Affect Survey, higher scores are indicative of more feelings of anger (irritability, frustration). Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of angry feelings and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of angry feelings. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Anger Survey

Description: This is a self-report measure for ages 8-17 as well as a parent-report measure for ages 3-12. It assesses angry mood and aggression (verbal and physical). The self-report version is a 6-item fixed-length form; the parent-report version for ages 8-12 uses a CAT format; and the parent-report version for ages 3-7 is a 9-item fixed-length form.

Scoring Process: For the self-report version, each item administered has a 5-point scale with options ranging from "never" to "almost always." The parent-report CAT for ages 8-12 utilizes a 4-point scale ranging from "almost never" to "almost always," while the parent-report version for ages 3-7 has a 3-point scale, ranging from "never or not true" to "often or very true." Each survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, which is then converted to an Uncorrected Standard Score (T-Score). An Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Anger Survey, higher scores are indicative of more child anger. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of anger and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of anger. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Fear is best characterized by symptoms of anxiety that reflect autonomic arousal and perceptions of threat. For adult self-report, fear is measured by the NIH Toolbox Fear-Affect Survey and the NIH Toolbox Fear-Somatic Arousal Survey; for child self-report, the NIH Toolbox Fear Survey is used; and for parent report, NIH Toolbox Fear-Over Anxious Survey and NIH Toolbox Fear-Separation Anxiety Survey are used for ages 3-7 and the NIH Toolbox Fear Survey for ages 8-12.

NIH Toolbox Fear-Affect Survey

Description: This self-report measure assesses fear and anxious misery for ages 18-85, using a CAT format. No versions for other ages are available.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, as is an Uncorrected Standard Score (T-Score).

Interpretation: For the NIH Toolbox Fear-Affect Survey, higher scores are indicative of more feelings of fear (fearfulness, panic). Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of fearful feelings and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of fearful feelings. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Fear-Somatic Arousal Survey

Description: This self-report measure assesses somatic symptoms related to arousal for ages 18-85, using a 6-item fixed-length form. No versions for other ages are available.

Scoring Process: Each item administered has a 5-point scale with options ranging from "not at all" to "extremely." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, which is then converted to an Uncorrected Standard Score (T-Score).

Interpretation: For the NIH Toolbox Fear-Somatic Arousal Survey, higher scores are indicative of more somatic arousal. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of somatic arousal and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of somatic arousal. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Fear Survey

Description: This is a self-report measure for ages 8-17 as well as a parent-report measure for ages 8-12. It assesses fear, anxious misery and hyperarousal. The self-report version is an 8-item fixed-length form; the parent-report version for ages 8-12 uses a CAT format.

Scoring Process: For the self-report version, each item administered has a 5-point scale with options ranging from "never" to "almost always." The parent-report CAT for ages 8-12 utilizes a 4-point scale ranging from "almost never" to "almost always." Each survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, as is an Uncorrected Standard Score (T-Score). An Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Fear Survey, higher scores are indicative of more child fear. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of fear and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of fear. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Fear-Over Anxious Survey

Description: This is a parent-report measure for ages 3-7, assessing fear, worry and hyperarousal. It is a 6-item fixed-length form.

Scoring Process: Each item administered has a 3-point scale, ranging from "never or not true" to "often or very true." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, which is then converted to an Uncorrected Standard Score (T-Score). An Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Fear-Over Anxious Survey, higher scores are indicative of more parent-reported child fear, worry and hyperarousal. Scores 1 SD or more below the mean (T ≤ 40)

suggest low levels of parent-reported child anxiety and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of parent-reported child anxiety. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Fear-Separation Anxiety Survey

Description: This is a parent-report measure for ages 3-7, assessing fear of being separated from home and from loved ones. It is a 7-item fixed-length form.

Scoring Process: Each item administered has a 3-point scale, ranging from "never or not true" to "often or very true." The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, which is then converted to an Uncorrected Standard Score (T-Score). An Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Fear-Separation Anxiety Survey, higher scores are indicative of more parent-reported child separation anxiety. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of parent-reported child separation anxiety and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of parent-reported child separation anxiety. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Sadness is distinguished by low levels of positive affect and comprised of symptoms that are primarily affective (poor mood) and cognitive (negative perceptions of self, the world and the future) indicators of depression. It is measured by the NIH Toolbox Sadness Survey.

NIH Toolbox Sadness Survey

Description: A self-report measure for ages 18-85 using a CAT format is available, as well as a self-report 8-item fixed-length form for ages 8-17; a parent-report measure for ages 8-12 using a CAT format; and a parent-report version for ages 3-7 that is a 7-item fixed-length form. Each survey version assesses negative mood, negative views of the self, and negative social cognition.

Scoring Process: For the self-report versions, each item administered has a 5-point scale with options ranging from "never" to "always" (adults) or "almost always" (ages 8-17). The parent-report CAT for ages 8-12 utilizes a 3-point scale ranging from "not true" to "true," and parent-report for ages 3-7 utilizes a 3-point scale ranging from "never or not true" to "often or very true." Each survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, as is an Uncorrected Standard Score (T-Score). For ages 3-17, an Age- and Gender-Corrected T-Score is also provided.

Interpretation: For the NIH Toolbox Sadness Survey, higher scores are indicative of more sadness. Scores 1 SD or more below the mean (T ≤ 40) suggest low levels of sadness and scores 1 SD or more above the mean (T ≥ 60) suggest high levels of sadness. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Supplemental Measures

Apathy is characterized by deficits in goal-directed behavior and decrements in goal-related thought content.

NIH Toolbox Apathy Survey

Description: This supplemental measure is a 7-item fixed-length self-report form for ages 18-85. No other versions are available.

Scoring Process: Each item administered has a 4-point scale with options ranging from “very true” to “not at all true.” The survey is scored using IRT methods. An IRT-derived theta score is generated for each participant, which is then converted to the Unadjusted Scale Score, in a column labeled T-Score on the Assessment Scores output file. No other normative data is available.

Interpretation: For the NIH Toolbox Apathy Survey, higher scores are indicative of more apathy. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of apathy and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of apathy. T-scores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Domain-Specific Life Satisfaction

Description: This supplemental self-report measure assesses feelings and attitudes about specific domains of one's life (e.g., family, health, work, leisure). A 13-item fixed-length form is used for adults, and 7-item fixed-length forms are used for ages 8-12 and 13-17, as well as for the parent-report versions for ages 3-7 and 8-12.

Scoring Process: Each item administered has a 5-point scale with options ranging from “not at all” to “very much.” Items are scored and reported individually but are not summed; therefore, no overall score is provided for this measure, and data will appear only on the data export.

Interpretation: For each item on the NIH Toolbox Domain-Specific Life Satisfaction Survey, higher scores are indicative of more life satisfaction in a given domain. Item responses can only be evaluated individually.

NIH Toolbox Maternal Relationship Survey

Description: For children, an important component of social support is their relationship with their parents. The NIH Toolbox Maternal Relationship Survey is a supplemental, self-report measure for children and adolescents ages 8-17. It is a 3-item fixed-length survey assessing the perceived quality of a child or adolescent's relationship with his/her mother in terms of an affective feeling of "closeness" and meaningful time spent together. It may be administered with the parallel Paternal Relationship Survey.

Scoring Process: Two items have a 6-point scale and one item has a 4-point scale, with all items gauging closeness with one’s mother. Items are scored and summed; the raw summed score is converted to Toolbox Age-Adjusted, Fully Adjusted and Unadjusted Scale Scores for PROs – Mean of 50, SD of 10 – as well as a national percentile rank that corresponds to the age-adjusted scale score. The Unadjusted Scale Score is provided in a column labeled T-Score on the Assessment Scores output file.

Interpretation: For the NIH Toolbox Maternal Relationship Survey, higher scores are indicative of better

perceived quality of the maternal relationship. Scores 1 SD or more below the mean (T 34 ≤ 40) suggest a poorer relationship and scores 1 SD or more above the mean (T ≥ 60) suggest a stronger relationship. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Paternal Relationship Survey

Description: For children, an important component of social support is their relationship with their parents. The NIH Toolbox Paternal Relationship Survey is a supplemental, self-report measure for children and adolescents ages 8-17. It is a 3-item fixed-length survey. It assesses the perceived quality of a child or adolescent's relationship with his/her father in terms of an affective feeling of "closeness" and meaningful time spent together. It may be administered with the parallel Maternal Relationship Survey.

Scoring Process: Two items have a 6-point scale and one item has a 4-point scale, with all items gauging closeness with one's father. Items are scored and summed; the raw summed score is converted to Toolbox Age-Adjusted, Fully Adjusted and Unadjusted Scale Scores for PROs – Mean of 50, SD of 10 – as well as a national percentile rank that corresponds to the age-adjusted scale score. The Unadjusted Scale Score is provided in a column labeled T-Score on the Assessment Scores output file.

Interpretation: For the NIH Toolbox Paternal Relationship Survey, higher scores are indicative of better perceived quality of the paternal relationship. Scores 1 SD or more below the mean (T ≤ 40) suggest a poorer relationship and scores 1 SD or more above the mean (T ≥ 60) suggest a stronger relationship. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Positive Parental Relationship Survey

Description: For children, an important component of social support is their relationship with their parents. As an analogue to the child and adolescent parental relationship surveys (maternal and paternal), the NIH Toolbox Positive Parental Relationship Survey is a supplemental, parent-report measure for children ages 3-12. It is a 5-item fixed-length survey that assesses the perceived positive qualities of the parent-child relationship from the perspective of the parent.

Scoring Process: Two items have a 4-point scale and three items use a 5-point scale. Items are scored and summed; the raw summed score is converted to Toolbox Age-Adjusted, Fully Adjusted and Unadjusted Scale Scores for PROs – Mean of 50, SD of 10 – as well as a national percentile rank that corresponds to the age-adjusted scale score. The Unadjusted Scale Score is provided in a column labeled T-Score on the Assessment Scores output file.

Interpretation: For the NIH Toolbox Positive Parental Relationship Survey, higher scores are indicative of more positive qualities in the parent-child relationship. Scores 1 SD or more below the mean (T ≤ 40) suggest a less positive relationship and scores 1 SD or more above the mean (T ≥ 60) suggest a more positive relationship. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Negative Parental Relationship Survey

Description: For children, an important component of social support is their relationship with their parents. As an analogue to the child and adolescent parental relationship surveys 35 (maternal and

paternal), the NIH Toolbox Negative Parental Relationship Survey is a supplemental, parent-report measure for children ages 3-12. It is a 4-item fixed-length survey that assesses the perceived negative qualities of the parent-child relationship from the perspective of the parent.

Scoring Process: Each item uses a 5-point scale. Items are scored and summed; the raw summed score is converted to Toolbox Age-Adjusted, Fully Adjusted and Unadjusted Scale Scores for PROs – Mean of 50, SD of 10 – as well as a national percentile rank that corresponds to the age-adjusted scale score. The Unadjusted Scale Score is provided in a column labeled T-Score on the Assessment Scores output file.

Interpretation: For the NIH Toolbox Negative Parental Relationship Survey, higher scores are indicative of more negative qualities in the parent-child relationship. Scores 1 SD or more below the mean ($T \leq 40$) suggest a less negative relationship and scores 1 SD or more above the mean ($T \geq 60$) suggest a more negative relationship. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

NIH Toolbox Sibling Rejection Survey

Description: As an analogue to the NIH Toolbox Perceived Rejection Survey, the NIH Toolbox Sibling Rejection Survey is a supplemental parent-report measure for children ages 3-12. It is a 9-item fixed-length form that assesses how often a child is left out, avoided or teased by siblings.

Scoring Process: Each item administered has a 5-point scale with options ranging from "never" to "always." Items are scored and summed; the raw summed score is converted to Toolbox Age-Adjusted, Fully Adjusted and Unadjusted Scale Scores for PROs – Mean of 50, SD of 10 – as well as a national percentile rank that corresponds to the age-adjusted scale score. The Unadjusted Scale Score is provided in a column labeled T-Score on the Assessment Scores output file.

Interpretation: For the NIH Toolbox Sibling Rejection Survey, higher scores are indicative of greater sibling rejection. Scores 1 SD or more below the mean ($T \leq 40$) suggest low levels of rejection and scores 1 SD or more above the mean ($T \geq 60$) suggest high levels of rejection. Tscores ≥ 60 may warrant heightened surveillance or concern. Additional work remains to examine the predictive and concurrent validity of these measures and to identify clinically meaningful thresholds.

Emotion Batteries and Summary Scores

Two NIH Toolbox Emotion Batteries are available: a self-report battery and a parent-report battery. The self-report battery is available for ages 8-85 and includes all age-specified measures in the Psychological Well-Being, Social Relationships, Stress and Self-Efficacy and Negative Affect domains. The parent-report battery is available for ages 3-12 and includes all age-specified measures in the Psychological Well-Being, Social Relationships, Stress and Self-Efficacy and Negative Affect domains.

In addition to scores for individual surveys as described above, each emotion battery provides summary scores, which allow for general interpretation/evaluation of overall emotional health and an even higher level of reliability than is possible with any individual survey score. To obtain a summary score, valid scores must be present for all component measures. The summary scores provided for each age group are as follows:

NIH Toolbox Emotion Summary Scores – Age 3 to 7 Parent-Report

Anxiety: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Fear-Over Anxious and Fear-Separation Anxiety scales.

Negative Psychosocial Functioning: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Sadness, Positive Peer Interaction, Social Withdrawal, Peer Rejection, and Empathic Behaviors. Prior to averaging, Positive Peer Interaction and Empathic Behaviors are reverse coded so that the direction is the same among all scales in the summary domain.

Psychological Well-Being: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Positive Affect, General Life Satisfaction, and Anger.

NIH Toolbox Emotion Summary Scores – Age 8 to 12 Parent-Report

Negative Peer Relations: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Positive Peer Interaction, Social Withdrawal, and Peer Rejection. Prior to averaging, Positive Peer Interaction is reverse coded so that the direction is the same among all scales in the summary domain.

Psychological Well-Being: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Positive Affect, General Life Satisfaction, Sadness, Anger, Fear, Self-Efficacy, and Perceived Stress.

NIH Toolbox Emotion Summary Scores – Age 8 to 12 Self-Report

Negative Affect: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Anger, Sadness, and Fear.

Social Satisfaction: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Friendship, Loneliness, and Emotional Support. Prior to averaging, Loneliness is reverse coded so that the direction is the same among all scales in the summary domain.

Negative Social Perception: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Perceived Hostility and Perceived Rejection.

Psychological Well-Being: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Positive Affect, General Life Satisfaction, and Self-Efficacy.

NIH Toolbox Emotion Summary Scores – Age 13 to 17 Self-Report

Negative Affect: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Anger, Sadness, and Fear.

Social Satisfaction: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Friendship, Loneliness, and Emotional Support. Prior to averaging, Loneliness is reverse coded so that the direction is the same among all scales in the summary domain.

Negative Social Perception: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Perceived Hostility and Perceived Rejection.

Psychological Well-Being: This age- and gender-corrected summary score is derived from a weighted average of age- and gender-corrected T-scores for Positive Affect, General Life Satisfaction, Self-Efficacy, and Perceived Stress.

NIH Toolbox Emotion Summary Scores – Age 18 to 85 Self-Report

Negative Affect: This summary score is derived from a weighted average of T-scores for Anger-Affect, Anger-Hostility, Sadness, Fear-Affect, and Perceived Stress.

Social Satisfaction: This summary score is derived from a weighted average of T-scores for Friendship, Loneliness, Emotional Support, Instrumental Support, and Perceived Rejection. Prior to averaging, Loneliness and Perceived Rejection are reverse coded so that the direction is the same among all scales in the summary domain.

Psychological Well-Being: This summary score is derived from a weighted average of T-scores for Positive Affect, General Life Satisfaction, and Meaning & Purpose.

Appendix A: Toolbox Standard Score to Percentile Conversion

Perf. Measure Standard Score	Percentile Rank	PRO Standard (T) Score	Perf. Measure Standard Score	Percentile Rank	PRO Standard (T) Score
140	99.6	77	99	47	49
139	99.5	76	98	45	48
138	99	75	97	42	48
137	99	75	96	40	47
136	99	74	95	37	47
135	99	73	94	34	46
134	99	73	93	32	45
133	99	72	92	30	45
132	98	71	91	27	44
131	98	71	90	25	43
130	98	70	89	23	43
129	97	69	88	21	42
128	97	69	87	19	41
127	96	68	86	18	41
126	96	67	85	16	40
125	95	67	84	14	39
124	95	66	83	13	38
123	94	65	82	12	38
122	93	65	81	10	37
121	92	64	80	9	37
120	91	63	79	8	36
119	90	63	78	7	35
118	88	62	77	6	35
117	87	61	76	5	34
116	86	61	75	5	33
115	84	60	74	4	33
114	82	59	73	4	32
113	81	59	72	3	31
112	79	58	71	3	31
111	77	57	70	2	30
110	75	57	69	2	29
109	73	56	68	2	28
108	70	55	67	1	28
107	68	55	66	1	27
106	66	54	65	1	27
105	63	53	64	1	26
104	61	53	63	1	25
103	58	52	62	1	25
102	55	51	61	0.5	24
101	53	51	60	0.4	23
100	50	50	59	0.3	23

Appendix B: Lookup Table for Words-in-Noise Test

	# Correct	Threshold		# Correct	Threshold
	0	26.0		20	10.0
	1	25.2		21	9.2
	2	24.4	MILD	22	8.4
	3	23.6		23	7.6
PROFOUND	4	22.8		24	6.8
	5	22.0		25	6.0
	6	21.2		26	5.2
	7	20.4		27	4.4
	8	19.6		28	3.6
	9	18.8		29	2.8
SEVERE	10	18.0	NORMAL	30	2.0
	11	17.2		31	1.2
	12	16.4		32	0.4
	13	15.6		33	-0.4
	14	14.8		34	-1.2
	15	14.0		35	-2.0
MODERATE	16	13.2			
	17	12.4			
	18	11.6			
	19	10.8			

Appendix C: LogMAR Score to Snellen Equivalency Table

LogMAR	Snellen	LogMAR	Snellen	LogMAR	Snellen
-0.3	20/10-0*	0.38	20/40-4	1.06	20/200-3
-0.28	20/10-1	0.4	20/50	1.08	20/200-4
-0.26	20/10-2	0.42	20/50-1	1.1	20/250-0
-0.24	20/10-3	0.44	20/50-2	1.12	20/250-1
-0.22	20/10-4	0.46	20/50-3	1.14	20/250-2
-0.2	20/12-0	0.48	20/50-4	1.16	20/250-3
-0.18	20/12-1	0.5	20/65	1.18	20/250-4
-0.16	20/12-2	0.52	20/65-1	1.2	20/320-0
-0.14	20/12-3	0.54	20/65-2	1.22	20/320-1
-0.12	20/12-4	0.56	20/65-3	1.24	20/320-2
-0.1	20/16-0	0.58	20/65-4	1.26	20/320-3
-0.08	20/16-1	0.6	20/80	1.28	20/320-4
-0.06	20/16-2	0.62	20/80-1	1.3	20/400-0
-0.04	20/16-3	0.64	20/80-2	1.32	20/400-1
-0.02	20/16-4	0.66	20/80-3	1.34	20/400-2
0	20/20	0.68	20/80-4	1.36	20/400-3
0.02	20/20-1	0.7	20/100	1.38	20/400-4
0.04	20/20 -2	0.72	20/100-1	1.4	20/500-0
0.06	20/20-3	0.74	20/100-2	1.42	20/500-1
0.08	20/20-4	0.76	20/100-3	1.44	20/500-2
0.1	20/25	0.78	20/100-4	1.46	20/500-3
0.12	20/25-1	0.8	20/125	1.48	20/500-4
0.14	20/25-2	0.82	20/125-1	1.5	20/640-0
0.16	20/25-3	0.84	20/125-2	1.52	20/640-1
0.18	20/25-4	0.86	20/125-3	1.54	20/640-2
0.2	20/30	0.88	20/125-4	1.56	20/640-3
0.22	20/30-1	0.9	20/160	1.58	20/640-4
0.24	20/30-2	0.92	20/160-1	1.6	20/640+
0.26	20/30-3	0.94	20/160-2		
0.28	20/30-4	0.96	20/160-3		
0.3	20/40	0.98	20/160-4		
0.32	20/40-1	1	20/200		
0.34	20/40-2	1.02	20/200-1		
0.36	20/40-3	1.04	20/200-2		

*The minus 0, 1, 2, 3, 4 designations in the Snellen column indicate missed optotypes. Therefore, 20/10-2 means the subject missed 2 letters at the 20/10 acuity level. The subject's LogMAR score can still be equated to 20/10.