Scoring information on each ABCD task

NIH Toolbox

For an overview of the dataset, there are 5 NIH tasks administered at all 3 timepoints and each 6 scores, an age corrected, uncorrected, raw, theta, computed and fully-corrected T-score (ABCD recommends using uncorrected scale or raw scores for longitudinal analyses)

According to the NIH manual:

Age corrected compares the score of the test-taker to those in the NIH Toolbox nationally representative normative sample **at the same age**

Fully corrected T-scores compare the score of the test-taker to those in the NIH Toolbox nationally representative normative sample, while **adjusting for demographic variables** (include age, gender, race/ethnicity, and educational attainment) collected during the NIH Toolbox national norming study. They provide this score to allows for comparison within a narrower grouping.

Uncorrected Standard Score: 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**. (It may also be of interest when monitoring performance over time).

For theta, all I found is: calculated for each participant; it represents the relative overall ability/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.

Computed: They only explain it for the Flanker test (the Flanker Computed score provides a way of gauging raw improvement or decline from Time 1 to Time 2)

The fluid intelligence and total composite score was not calculated for every wave because it includes 2 tests that were not followed up, so it cannot be used. That should also make the crystallized score not relevant?

Raw scores not available for picture vocabulary and reading

The Flanker task (inhibition/fluid intelligence) used the Inquisit system from Millisecond (designed to mimic the NIH Toolbox Flanker task as closely as possible) for remote assessments in the 2-year and 4-year follow-up protocols. Despite covid, not all assessments happened remote/hybrid for the last 2 waves, so some subjects have one version and some the other. Should I keep the test?

Also, the Little man task matches the time points of the NIH toolbox (mental rotation) but in the baseline assessment a customized program designed by ABCD was used, whereas the 2-year and 4-year follow-up assessments a task presented in the Inquisit system from Millisecond was used (regardless of remote/in person visit).

Κρατάω: ID, eventname, raw scores. Ιδανικά προσθέτω στήλη με visit type και πιο ιδανικά τη στήλη με wm (ελέγχω αν έχω τη σωστή)

1. *NIH Toolbox Picture Vocabulary Test (TPVT): Measure of receptive vocabulary and crystallised intelligence, administered in a computerized adaptive format.*

*Scoring Process: Item Response Theory (IRT) is used to score the TPVT.*

**Theta scores**: calculated for each participant; it represents the relative overall ability/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.

To interpret individual performance, one can evaluate all three types of scores.

**Age-Corrected**: 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)

**Uncorrected Standard Scores**: 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. (but matched in which domains?)

**Fully Corrected T scores**: 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.

**Raw scores**: 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.

-associated Percentiles can be found in Appendix A.

1. NIH Toolbox Oral Reading Recognition Test (Reading): the participant is asked to read and pronounce letters and words as accurately as possible.

Scoring: IRT

To interpret individual performance, one can evaluate all three types of standard scores plus the percentiles; 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.

1. Flanker Inhibitory Control and Attention Test (Flanker): 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

Scoring: 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 explained, Reaction time vector explained.

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).

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.

1. Picture Sequence Memory Test (PSMT): 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.

Scoring: 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.

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.

1. *Pattern Comparison Processing Speed Test: processing speed.*

*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****. 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.*

FLANKER EXPLANATION

Requires the participant to focus on a given stimulus while inhibiting attention to stimuli

3-7: fish, 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.

8-85: arrows. Twenty trials are conducted for ages 8-85;

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: • Flanker o Fish: 20 Points o Arrows: 20 Points

I think: uncorrected scores are “richer” ONLY for people with high accuracy, ONLY for the uncorrected version. The difference between scores is explained because everyone in the raw is scored the same way, based only on their accuracy, but in the raw, some people (the high scoring ones) are also scored based on their response time. So, for the people who score highly on accuracy, their uncorrected scores can/will be different from the raw, they will have more variety (for a raw score of 20, so 20/20 accuracy, you can have slower of faster response times, which will be reflected in your uncorrected score but not in your raw score). So the uncorrected score is not a simple transformation of the raw. If you score less than 80% in accuracy, your raw and uncorrected scores will reflect the same thing, accuracy, so one child will have identical uncorrected scores to another if we only know its raw score.

* For a given **raw score of 20 (perfect accuracy)**, uncorrected scores can vary significantly depending on how fast or slow the participant responded.

Largest discrepancies between estimation methods (standard-SAM)

# flanker\_intercept-pattern\_intercept: 0.63 0.49

# flanker\_slope-picture\_slope: 0.23 0.04

# picvocab\_slope-reading\_slope: 0.55 0.93

# flanker\_slope-pattern\_slope: 0.78 0.36

# picvocab\_slope-reading\_slope: 0.55 0.93

# picvocab\_slope-picture\_slope: 0.61 0.13

# picture\_intercept-picture\_slope: -0.09 -0.41

# pattern\_intercept-pattern\_slope: -0.13 -0.40