National Institutes of Health Toolbox Cognitive Battery Supplemental Administrator's Manual for Intellectual and Developmental Disabilities
A Guide on Administration and Scoring Standards

Forrest McKenzie, Andrea Drayton, Rebecca Shields, M.S., Aaron Kaat, Ph.D., Jeanine Coleman, Ph.D., Talia Thompson, Ph.D., Stephanie Sansone, Ph.D., Karen Riley, Ph.D., Elizabeth Berry-Kravis, M.D., Ph.D., Richard Gershon, Ph.D., David Hessl, Ph.D.

For the most updated version of this document, please see the NIH Toolbox support website:
https://nihtoolbox.desk.com/customer/portal/articles/2981718

For questions, please contact Dr. David Hessl

Translational Psychophysiology and Assessment Laboratory, MIND Institute, UC Davis, 2825 50th St. Sacramento CA 95817, 916-703-0249, drhessl@ucdavis.edu

Version 1.0
8/6/19

T-PAL
Translational Psychophysiology
and Assessment Laboratory

Table of Contents

A. Introduction

This manual was created to be supplementary material to the *National Institute of Health (NIH) Toolbox for the Assessment of Neurological and Behavioral Function*® *Administrator's Manual*[1] for use when testing examinees with intellectual and developmental disabilities (IDDs). Our work from a large-scale study to empirically validate the NIH Toolbox Cognitive Battery (NIHTB-CB) in intellectual disabilities has led to the creation of standards on test administration specific to this population, in order to address accessibility and feasibility issues, scoring dilemmas, and the test environment. If administering the NIHTB-CB to individuals with IDD, we recommend following all procedures in the original manual, as well as the additional information presented in this addendum. This manual does not serve to replace the standard manual, but rather to be supporting material to aid examiners specifically working with the IDD population.

B. Examinee Eligibility

The National Institute of Health Toolbox Cognitive Battery was developed for examinees ages 3-85. However, individuals with IDD have a mental age equivalent that is substantially lower than their chronological age. So, even though individuals with IDD may have a chronological age that falls within the range outlined by the NIHTB-CB, in order to begin testing at an appropriate level of difficulty it is strongly advised to use an examinee's estimated mental age rather than chronological age when deciding if the CB is an appropriate assessment. Based on extensive feasibility studies, we recommend that the NIHTB-CB is used for examinees with a mental age of at least 3 years, 0 months, although feasibility can be quite variable above this level depending on the NIHTB-CB test. Additionally, it is important to use the mental age rather than chronological age when setting up the assessment (see pages 3 and 4).

Prior to setting up and administering assessments in the Toolbox application, the examiner will need to determine the examinee's approximate mental age. One type of score provided by many IQ tests is the mental age equivalent. Our group uses the Stanford Binet Intelligence Scales, 5th Edition,[2] to determine the examinee's full scale mental age prior to administering the NIHTB-CB. If the mental age is unavailable or unable to be determined by the Stanford Binet or a different IQ test, we recommend using available clinical and educational records, prior cognitive tests, adaptive behavior assessments, or clinical judgment of developmental level based on common cognitive milestones to estimate the examinee's current mental age.

C. Setting Up the Toolbox Cognitive Battery

When entering the examinee's demographic information on the examinee registration screen enter the actual date of birth if desired (such as for use of age-corrected scores or T-scores), along with other demographic information that you would like to include. After entering demographic information, you can begin creating the assessment by selecting the tests you would like to administer. As previously described, when selecting test versions use the examinee's mental age. For example, an examinee with Down syndrome with a chronological age of 11 years, 2 months would not be given the test versions that correspond with her chronological age of 11 years. Instead, her mental age would be determined and used to select the appropriate test versions. So, if her mental age was 5 years, 4 months, then the examiner would select the test versions for 5 year olds. For example, if selecting Pattern Comparison, the examiner would select "Pattern Comparison Processing Speed Ages 3-6."

C.1 Education Override and its Calculation

Education Override is a feature on the examinee registration screen. It allows the examiner to adjust the starting point for the computer adaptive tests that are part of the NIHTB-CB--the NIH Toolbox Picture Vocabulary Test and the NIH Toolbox Oral Reading Recognition Test. If an education override is inputted, the NIHTB-CB start point is based on the education override for these two tests, not the chronological age or current grade. If an education override is not entered prior to assessment, the NIHTB-CB will instead base its starting points on the examinee's actual grade level. For testing individuals with IDD, we recommend using education override so that the test does not begin with items that are too far above the examinee's ability level. Note that the education override option only becomes available once the relevant tests are added to the assessment. Thus, after creating the assessment by selecting the tests that you want to include, you will need to go back to the examinee registration section and enter the education override before beginning the assessment. (If you are not using the Picture Vocabulary Test or the Oral Reading Recognition Test, there is no option or need for education override.)

To calculate the grade estimate used in education override, an estimated mental age is needed. As described earlier, the mental age can be obtained in several ways. If there is no documentation for mental age the examiner may estimate using clinical judgment. Below is the calculation that we recommend for determining the grade estimate for education override:

> **Mental Age Estimate rounded down to the year – 5 = Grade Estimate**

For example, an examinee whose current mental age estimate is 3 or 4 years would have a negative converted grade estimate. The examiner would select "Preschool" in the education override feature, as subtracting 5 from 3 or 4 yields a negative number. If the examinee has a mental age of 7 years, the converted grade estimate would be 2 because subtracting 5 from 7 yields 2. The examiner would select 2nd grade in the education override feature. Table 1 on the following page presents some examples of possible mental ages and their corresponding converted grade estimate and education override selection.

Table 1: Education Override Reference Table		
Mental Age Estimate rounded down to the year	**Converted grade estimate**	**Corresponding education override selection**
3 years, 4 years	-2, -1	Preschool
5 years	0	Kindergarten
6 years	1	1st grade
7 years	2	2nd grade

Note: If the examinee will be testing with the NIHTB-CB in the future, it is important to remember to update education override in the registration screen prior to retesting. Be sure to use their <u>current</u> mental age at each assessment.

C.2 Developmental Extension

Developmental extensions (DEXTs) are <u>experimental</u> versions of two tests--Flanker Inhibitory Control and Attention (Flanker) and Dimensional Change Card Sort (DCCS). The DEXT versions are intended to extend the range of these tests downward for lower functioning and/or younger individuals, but note that the scoring of these versions is currently under investigation. DEXT allows an examinee who does not pass the original practice items or who does pass practice but performs below a certain threshold on test items to then receive simpler items with easier, step-by-step instructions and practice. If the examiner administers the test version with developmental extension, then the NIHTB-CB will automatically present DEXT items if they are needed. This allows examinees to learn the task in a more accessible way if they struggle with the normal items due to poor understanding or other factors.

The following NIHTB-CB tests currently have a DEXT version: NIH Toolbox Flanker Inhibitory Control and Attention Test Ages 3-7 with Developmental Extension and NIH Toolbox Dimensional Change Card Sort Ages 3-7 with Developmental Extension. See the NIH Toolbox App Administrator's Manual for more detail on DEXT, including the criteria for receiving DEXT items.

Although DEXT items address the need in this population for test items at a lower range, there are some limitations with using the DEXT versions in their current experimental form. First, the scores obtained on DEXT items are not the same type of scores obtained on the normal test. DEXT scores are the raw number and percent correct out of total DEXT items. Thus, DEXT scores cannot be combined or compared with the validated computed or standard scores from these tests. Scores using these extension versions should currently be used only for research purposes and interpreted with caution.

D. Administering the NIHTB-CB

When administering the NIHTB-CB it is important that it is done in an environment that is as free from distractions as possible and is conducive to an ideal testing atmosphere. Testing in a private setting and minimizing disturbances will help the examinee focus on the tasks. Sitting directly next to the examinee allows for easier redirection if distraction occurs. It is also important to sit in view of the iPad as there are directions to be read and responses to be reported by the examiner. Furthermore, sitting next to the examinee and far enough away from exits will lessen the likelihood of escape behaviors while testing.

D.1 A Model for Accommodations

Accommodations are an important part of the assessment process. Our team has developed a model for implementing appropriate accommodations that follow a typical assessment cycle of planning, administration, evaluation and reporting. Each section, as seen below, has components and questions to guide the examiner through the process so that the individual with IDD has an equitable opportunity to participate in the assessment process and maximize their potential to demonstrate abilities. In order to achieve equity, assessment examiners need to address "target skills" that are the intended construct. "Target skills" refer to the construct that an assessment intends to measure, while "access skills" are those abilities – such as joint attention, fine motor skills, and language ability – that are required before an individual can demonstrate his or her understanding of testing conditions, content, and instructions during a standardized assessment.[3] For example, if a test addresses attention, the examiner may need to make some accommodations, but should not provide attentional scaffolding (e.g., "look at the screen", or "pay attention!") beyond the standardized instructions that could invalidate scores arising from testing. Please see Thompson and colleagues (2018) for more detail.[3]

Additionally, it is important to consider the assessment environment, and the influences on the individual and their family. Many individuals experience anxiety prior to and during the session, due to logistics (i.e., parking, walking into the clinic), prior negative experiences, change in their routine, or meeting new people. To minimize these factors, the team can reduce unfamiliar and unpredictable experiences before, during, and after the assessment sessions. Pre-visit strategies include sending pictures, video, or Social Stories[4] of staff summarizing the experience ahead of the appointment; individualizing greetings to the individual and family by modulating tone of voice; limiting the number of people during the greeting; and allowing time for rapport building and exploring the environment. The team can indicate verbally and visually the location of services and facilities available for use, such as restroom, elevators, and kitchens/lounges. Often, an extended "warm-up" period is necessary to ease anxiety in examinees and transition into the assessment environment.

D.2 Reinforcement Inventory and Token Board

Particularly when working with examinees with IDD, motivation to complete testing with good effort may be a problem. Using reinforcements and rewards can be especially helpful to maintain examinee motivation. Before administering the NIHTB-CB it is recommended that information about the examinee's preferences regarding snacks and rewards is collected. The Reinforcement Inventory form may be used for this purpose and can be found at the end of this manual. Knowing about the examinee's likes and dislikes will help the examiner to select items

that would motivate the examinee during testing. This form can be sent via email and reviewed prior to testing with enough time to obtain items that might be helpful.

The token board is a visual representation of the NIHTB-CB tests. Once the examinee completes a test, he can cross off or remove the matching picture. Involving the individual in this way actively informs him of how many tests he has left and regularly motivates him to continue with the assessment. An example of a token board used while testing is shown below and a printable size can also be found at the end of this manual.
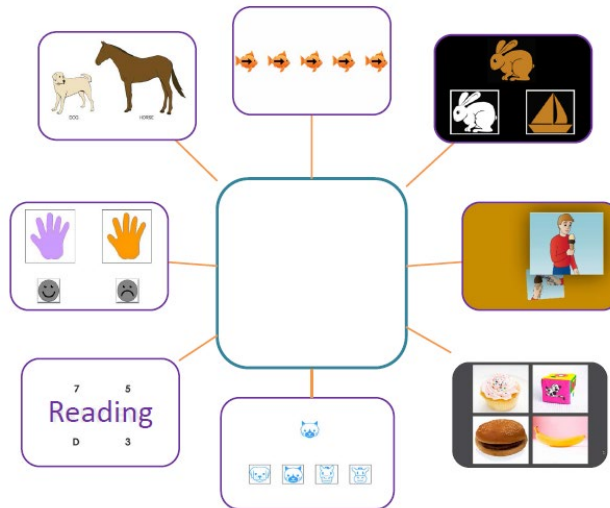

Figure 1. Example of the token board used during administration the NIHTB-CB. The examiner can write or place the reward in the blank box in the middle.

The examiner can allow the examinee to choose a prize that will be given after the completion of the NIHTB-CB. The name of the prize, or the prize itself if small enough, can be written or placed at the center of the token board for reference. Additionally, if the individual needs more incentive to continue testing, the examiner may give him or her food or small items such as stickers after each test. Food may also be given during certain breaks within tests so long as the examiner follows the correct reinforcement protocol for each test as outlined in the chart below. However, this should be a last resort and only according to the guidelines for each specific test listed in the Reinforcement Guidelines table (below).

During the visit, examiners can clearly communicate the timing of breaks and indicate the expected duration of the assessment. If needed, caregivers or aids may be incorporated into the testing room to provide a stabilizing and motivating influence for the examinee, but it is preferred to first try testing without a third person present. If present, the caregiver may position themselves out of direct view, and should be told not to participate in the testing process. For some individuals, session length may require adjustment, as they may work best with multiple shorter sessions. Visual checklists, fidget items, and the opportunity for movement breaks are important accommodations in the testing process. Age is also a consideration, and younger examinees may require additional equipment (e.g., footstools or lower tables) in order to fully access the session. Supportive seating is important, as many examinees may have low muscle

tone or hyperactivity - a secure, developmentally appropriate seat tends to minimize these problems. Remaining flexible and providing choices are critical to reducing anxiety and maximizing examinees' access to the assessment. Use of humor can also help with anxiety in many examinees. Finally, examiner preparation is very important. Examiners should be intimately familiar with the NIHTB-CB. Examiners should be well practiced with administering the NIHTB-CB by first practicing administration with volunteers so that all of their attention and efforts can be directed to the examinee to maximize performance, rather than being distracted or slowed down by lack of familiarity with the system or the participant population. We recommend that a clinically trained assessment expert with experience evaluating people with IDD observe several practice testing sessions and provide feedback to newer examiners before "real" assessments are done. This supervisor can use their professional judgement to determine when the examiner is prepared adequately.

D.3 Reinforcement Guidelines

Your experience from working with the examinee while administering the Stanford Binet or equivalent and the practice items from the NIHTB-CB will be a good indicator of the level and frequency of redirection and reinforcement necessary to keep the individual engaged and motivated while administering the test items. An important note is that reinforcements should not be dependent on responding correctly to an item; they should be focused on the individual's behavior and staying on task. The purpose of providing reinforcement is to increase and maintain motivation during testing.

The most frequent reinforcement timing is listed in Table 2 for each specific test. Reinforcement may be given less frequently than recommended in the table on the following page, but not more frequently. For example, if the examinee is highly motivated and cooperative during a task that allows trial-based reinforcement the assessor may choose to give reinforcement after completion of the whole task.

The following describes the difference between task-based and trial-based reinforcement:

1) Task-based (less frequent schedule) involves offering a small reward after the completion of every task and a larger reward at the end of all tasks.
2) Trial-based (more frequent schedule) involves offering a small reward after every trial or ratio of trials (e.g. reinforcement provided after completion of 5 items/trials).

It is important to note that tests allowing task-based reinforcement (rather than trial-based) do so because of a time component intrinsic to the test. Meaning, reinforcement during the task may slow down the examinee's responses, which will negatively impact their score. So, for these tasks it is best to reinforce at the end or when there is a break within the task.

Table 2: Reinforcement Guidelines	
Dimensional Card Change Sort Test	Task-based; reinforcement can be given after completion of all items
Flanker Inhibitory Control and Attention Test	Task-based; reinforcement can be given after completion of *fish set* and/or after completion of the *arrows set*
Picture Sequence Memory Test	Sequence-based; similar to the trial-based schedule, reinforcement may be given at the end of each completed sequence
Picture Vocabulary Test	Trial-based; reinforcement can be given after a certain number (decided by examiner) have been completed by examinee
Oral Reading Recognition Test	Trial-based; reinforcement can be given after a certain number (decided my examiner) have been completed by examinee
Pattern Comparison Processing Speed Test	Task-based; reinforcement can be given after completion of all items
List Sorting Working Memory Test	List-based; similar to the trial-based schedule, reinforcement may be given after the examinee completes their entire response and before starting the next list
Speeded Matching	Task-based; reinforcement can be given after completion of all items

D.4 NIHTB-CB Administration Form

During the administration of the NIHTB-CB the examiner should record notes on the individual's performance. This is important in order to be able to reference testing environment and examinee performance notes later. It is also important when deciding if an individual's score on each test should be considered a "valid" score--that is, if their score represents their performance on the construct that the test is intended to measure. The notes can include if practice was passed, if DEXT items were administered, and if the test was valid. It also allows the examiner to take notes on the testing situation with the individual and any issues during the administration.

The NIHTB-CB Administration Form includes fields for these notes and descriptions of common validity issues (see appendix). The Administration Form is based on our large-scale study of the NIHTB-CB in IDD. The form is useful for testing any examinee, but is especially important in IDD. Of course, it is up to the examiner or researcher to decide how to document notes and what scores they consider valid, but the Administration Form is the standard form that we have developed through extensive assessment in IDD.

If planning to randomize the subtest order, there is a blank column to document the order. The Administration Form documents whether each test score should be considered "valid" with space to record reasons for invalidity, with checkboxes of common reasons that we have encountered and space to write other reasons. In addition to selecting one or more checkboxes when a test is invalid or questionable, we recommend always expanding on this with more information in the notes space. The last column includes space for additional notes about each test if needed. It also includes a checkbox to note if the examinee's understanding of the task was questionable. If the examinee passes practice this is not necessarily a reason to consider the score invalid, but we find this useful to document.

The last page of the form describes common issues seen in this population, which may result in the subtest being invalid. It is useful to have this information when interpreting and analyzing the scores.

E. NIHTB-CB Subtest Adaptations

Some tests are more difficult for individuals with IDD to understand, so we recommend modifications to some of the instructions. These tests include the NIH Toolbox Touch Screen Tutorial and NIH Toolbox Picture Vocabulary Test. Additionally, we recommend an adjusted administration protocol on Picture Sequence Memory (Picture Sequence Memory) to address potential floor and ceiling problems.

E.1 NIHTB-CB Touch Screen Tutorial

The Touch Screen Tutorial helps to orient the examinee with using the iPad. Although it is not typically a requirement, we recommend using the Touch Screen Tutorial at the beginning of the assessment for all examinees with IDD. We have modified the instructions for the Touch Screen Tutorial in order to also introduce "Home Base" before beginning actual testing. As described in the NIH Toolbox App Administrator's Manual, home base is used in portions of Flanker and DCCS if the examinee reaches that level of the test. The purpose of home base is to standardize reaction time, as portions of these tests take reaction time into account for scoring.

Examinees with IDD benefit from an introduction to home base prior to testing. This way they can more thoroughly become familiar with home base rather than only receiving the standard brief introduction in the middle of the test. Thus, we have modified the instructions for the Touch Screen Tutorial so that the tutorial can simultaneously serve as an introduction to home base. The modified instructions are on the following page. If the examinee does not understand or demonstrate proper use of home base during Touch Screen Tutorial, we recommend not using home base during tests (even if the examinee advances to the point where home base would be introduced, skip over those instructions about home base during the test).

T-PAL
Translational Psychophysiology
and Assessment Laboratory

Added Instructions for NIHTB-CB Touch Screen Tutorial

On the touch screen tutorial start page say the following: **Let's start by learning about home base. This is your home base. Put your finger on the home base and wait for the next picture. In this game, put your finger on home base after you touch the screen.** *Point to the home base; then touches and holds the button to continue*

Touch the flower on the screen. If it appears that the examinee would have difficulty touching the screen and then putting their finger back on home base, the examiner may use this item as a teaching item to demonstrate how to use home base and say, **See first I touch the flower** *examiner touches flower* **then I place my finger back on home base.**

Good! Now we are going to do some more. Remember to start with your finger on home base.

Now you touch the star.

If the examinee responds correctly, by touching the star and returning to home base then say, **Good! Let's do one more.**

If they do not respond correctly remind them how to use home base.

Now touch the apple.

If they do not respond correctly you may discontinue the use of home base for the remaining tasks.

E.2 NIHTB-CB Dimensional Card Change Sort Test (DCCS)

While completing the practice items for this test, the iPad will tell the examinee if their answer is correct or incorrect. But, during the test items, the iPad no longer gives this feedback. Some individuals may fixate on this and might begin to change their answers or not perform as well. The added instructions for addressing this are below. These can be used if the examiner anticipates that they will be helpful.

Added Instructions for NIHTB-CB Dimensional Card Change Sort

To be read before the test items begin: **This time the iPad will be quiet and will not tell you if your answer is right or wrong.**

E.3 NIHTB-CB Picture Vocabulary Test

Picture Vocabulary has two sets of instructions depending on the education override that is used. For examinees with an education override of preschool, kindergarten, or first grade, short instructions are presented to the examiner to read to the examinee. Alternatively, if the examinee has an education override of at least second grade, longer, more detailed instructions are presented.

It is our recommendation that for all examinees in the IDD population, regardless of their education override, the short instructions be used as they are more easily understood and not cumbersome; the shorter length provides less opportunity for distraction. The short instructions are below and can be printed for use during administration.

Short Version of Instructions for NIHTB-CB Picture Vocabulary Test
Read these instructions in place of the on-screen instructions: **Let's look at some pictures. You'll hear a word and see four pictures on the screen. Touch the picture that means the same as the word that was said. Some words will be easy and some will be harder. Answer as best as you can. If you need to hear a word again, I can play it again.**   **If you think you made a mistake and want to change your last answer, I can bring back that word and pictures. Are you ready to play?**   *After practice items*: **Now, let's try some more. Remember, you will hear a voice say a word and then you will see four pictures. One of the pictures will show what the word means. Touch that picture and you will hear a new word and see four more pictures. This time the iPad will be quiet and will not tell you if your answer is right or wrong. Any questions?**

E.4 NIHTB-CB Flanker Inhibitory Control and Attention Test

As previously described in the NIH Dimensional Card Change Sort section, the NIH Flanker Inhibitory Control and Attention Test also may need the added instructions clarifying that the iPad will not give feedback during the test items. These can be used if the examiner anticipates that they will be helpful.

Added Instructions for NIHTB-CB Flanker Inhibitory Control and Attention Test
To be read before the test items begin: **This time the iPad will be quiet and will not tell you if your answer is right or wrong.**

E.5 NIHTB-CB List Sorting Working Memory Test

As previously described in other tests, the NIH List Sorting Working Memory Test may also need the added instructions clarifying that the iPad will not give feedback during the test items. These can be used if the examiner anticipates that they will be helpful.

Added Instructions for NIHTB-CB List Sorting Working Memory Test
To be read before the test items begin: **This time the iPad will be quiet and will not tell you if your answer is right or wrong.**

E.6 NIHTB-CB Picture Sequence Memory Test

In our validation study in IDD, we have found a common problem of floor and ceiling issues on Picture Sequence Memory--specifically, examinees (with a mental age between 3 and 4 years) receiving the Age 3-4 version sometimes score at ceiling, and examinees (with a mental age between 5 and 6 years) receiving the Age 5-6 version sometimes score at the floor. Because there are multiple age versions, we recommend that the floor or ceiling score should not be used, and instead the examinee should additionally be tested on the next highest or lowest version (if available). Their score on that additional version should replace their floor or ceiling score, and the examiner should note which version was used. Note that floor and ceiling scores on Picture Sequence Memory can be harder to monitor for and to check, as the lowest and highest possible theta score depends on the test version. It is also difficult during testing to monitor if the examinee is getting everything correct or incorrect.

Since we have found this problem specifically with the Age 3-4 and Age 5-6 versions, we have adapted our administration protocol to always give both of these versions for examinees with a

mental age between 3 and 6 years, as a current solution. Our administration guidelines below detail this procedure. This is to provide greater assurance that a valid score will be obtained from testing. These two versions should be administered back-to-back. See the scoring section of this manual for further discussion of Picture Sequence Memory scores and the impact of this issue on composite scores.

Adapted Administration Guidelines for NIHTB-CB Picture Sequence Memory

---

When setting up the assessment for an examinee with a mental age between 3 years, 0 months and 6 years, 11 months: select both Picture Sequence Memory Age 3-4 and Picture Sequence Memory Age 5-6. Picture Sequence Memory Age 3-4 should be administered <u>first</u>.

Administer Picture Sequence Memory Age 3-4 with the standard directions. If the examiner was able to observe that the examinee did not get every item correct, then the Age 5-6 version does not need to be administered and can be manually skipped.

If the examinee placed every item correctly on Age 3-4, then the Age 5-6 version should be administered. If the examiner is unsure if all items were placed correctly, then the Age 5-6 version should be administered. The examiner should not read the instructions again or have the examinee complete the practice. We recommend that the examiner briefly mention that they are preparing the next part, mute the volume, pick up the iPad, and quickly go through teaching and practice sequences to advance the test. After the two practice sequences, the examiner should unmute the volume, return the iPad to its place in front of the examinee, and have the examinee continue with the test sequences of Age 5-6.

If the examiner or researcher does not want to give the dual administration by default, another option would be to give the version corresponding with mental age. Then, after testing is complete, the examiner can check the report to see if the score was at the floor or ceiling for that version. The examiner can check raw scores to easily check for lowest or highest possible scores; a raw score of 0 would be the lowest possible score (though not accounting for random chance of correct placements), and the highest possible raw score would be 10, 16, or 31 on Age 3-4, 5-6, or 7/8+ versions, respectively. In the case of a floor or ceiling score, the examiner would create a second administration with only Picture Sequence Memory and select the next older or younger version to administer immediately. For example, if an examinee has a score at the floor of Picture Sequence Memory Age 5-6, they should then receive Picture Sequence Memory Age 3-4. If an examinee receives a floored score on Picture Sequence Memory Age 3-4, there is no younger version available, so they would be done with Picture Sequence Memory. If an examinee has a score at the ceiling of Picture Sequence Memory Age 3-4, they should then receive Picture Sequence Memory Age 5-6. See the scoring section for further discussion of Picture Sequence Memory scores and the impact of this issue on composite scores.

F. Common Administration Issues

During the validation of the NIHTB-CB in IDD there have been several issues that repeatedly surface during administration and these are addressed in this section.

A common issue in the NIH Pattern Comparison test is that during the test items some examinees begin to alternate their answers back and forth from "Yes" to "No" without consideration of the stimuli. Even though the examinee may have passed practice and have an understanding of how to complete the task, they may revert to this alternating response pattern. Examiners will need to pay attention to the response behavior of some examinees to detect this during the testing process. In this case, the test would be invalid and should be noted on the Administration Form. The examiner should be relatively sure that this invalid pattern is occurring in order to deem the test invalid. (Note: If an examinee is simply responding in this alternating pattern, they may do it quickly and obtain a reasonably high, but invalid score due to the fact that each response has a 50% correct chance.) This issue is a primary reason that we have developed a new processing speed test, Speeded Matching (described below) which is currently being psychometrically evaluated.

A common issue in List Sorting Working Memory is that the examinee forgets the rule by which they need to repeat back the items. In the 2-list portion, it is allowable to restate the instructions "tell me the foods first, then the animals," as described in the NIH Toolbox Administrator's Manual. This reminder is often needed with this population.

G. Experimental Subtest Development

Along with DEXT versions (described earlier), the following Experimental Subtests are also beneficial for assessing individuals with IDD. However, note that because these are new tests that were not included in the original NIHTB-CB standardization, no standard scores are yet available.

G.1 Speeded Matching

Due to the Pattern Comparison Processing Speed Test being one of the least feasible tests for examinees with IDD, a new processing speed test has been developed as an alternative. Speeded Matching is a processing speed test and is available in the app as an experimental version. In Speeded Matching, the examinee is presented with one animal face centered at the top of the iPad screen and four animal faces on the bottom of the screen. The examinee searches amongst the four pictures to find the same picture presented above as quickly as possible, and then the next test item appears. This test design prevents the alternating response pattern some examinees exhibit during Pattern Comparison Processing Speed as there are more than two answer choices, and the task of matching is straightforward and familiar. Speeded Matching is performing very well; it is highly feasible in examinees with IDD, and future psychometric studies will present reliability and validity results. Note that only Speeded Matching raw scores are available; standard, norm-referenced scores are not yet available, and as such this test does not factor into any CB composite scores.

G.2 List Sort Working Memory Test Age 3-6 with Extended Instructions

Because List Sorting Working Memory was also one of the least feasible in our population, extended instructions have been tested to improve feasibility by more clearly teaching what is being asked of the examinee. In the normal Age 3-6 practice, the examinee must state the animals in size order one at a time, but rearranging items is not required, as they are already sequenced when they appear. The extended instructions build on this by adding additional practice items in which the examinee must name more than one item at a time and must rearrange the animals rather than repeat the (already ordered) sequence. Thus the practice resembles the real task. These extended instructions have been implemented in the List Sorting Working Memory Age 3-6 with Extended Instructions Experimental Version in the app.

H. Critical Considerations for Scoring and Analysis

The NIH Toolbox app generates several types of scores for each Toolbox test and composite. See the Health Measures website[5] and the *NIH Toolbox Scoring and Interpretation Guide*[6] for scoring information. Here we briefly describe each type of score. We then go into detail about important considerations when working with IDD.

For the standardized (non-experimental) tests, there are raw scores, computed scores, theta scores, and standard test scores. There are also composite standard scores. The raw scores are simply raw numbers of items or points correct. Computed scores are available on some of the tests; on Flanker and DCCS, for example, the computed scores are calculated from reaction time and accuracy combined. On Picture Sequence Memory, Picture Vocabulary, and Oral Reading, Item Response Theory-based scoring is used and theta scores are provided; the theta score represents the examinee's performance in comparison to normative scores. Uncorrected Standard Scores represent performance relative to the general population. Because they are not adjusted for age, these can reflect true change over time. Age-Corrected Standard Scores represent performance relative to the same age range in the general population. Fully Corrected T-Scores are further adjusted based on gender, education, and race/ethnicity.

The Age-Corrected Standard Scores and Fully Corrected T-Scores require special caution for this population because these scores are currently winsorized; the floor is imposed to just below three SDs below the mean (similarly, the ceiling is imposed to just above three SDs above the mean). This floor (54 on Age-Corrected standard scores and 19 on T-Scores) is a problem when evaluating individuals with IDD, as their scores are often below this cut-off. The NIH Toolbox has a SAS rescoring script available online, which can be used to generate an examinee's non-winsorized scores if the Age-Corrected or T-Scores are desired.

The NIH Toolbox has four types of composites: Fluid (scored from DCCS, Flanker, List Sorting Working Memory, Picture Sequence Memory, and Pattern Comparison Processing Speed), Crystallized (from Picture Vocabulary and Oral Reading), Cognitive Function (scored from all seven tests), and Early Childhood (scored from DCCS, Flanker, Picture Sequence Memory, and Picture Vocabulary). Reference can be made to Cassaletto et al. for composite score details and formulae.[7]

For List Sorting Working Memory Age 3-6 with Extended Instructions and Speeded Matching, standard scores are not currently available. These tests provide raw scores. For Flanker DEXT and DCCS DEXT, in addition to the normal test scores from performance on the normal portion of the test, if an examinee receives DEXT items, they will receive raw DEXT scores (number correct, number attempted, and percent correct out of those attempted). Since the DEXT scores differ from the normal test scores, they cannot be directly compared to or combined with normal test scores. New versions of DEXT tests are under consideration. With the current versions, the examiner must be aware that DEXT scores are on a different scale than the normal test scores.

There are several considerations that are especially important regarding scoring for this population. First, as mentioned above regarding the Administration Form, monitoring test session validity is critical in examinees with IDD. After testing, the Administration Form should be used to assist in determining if the score on each test is valid. This process must be done before reporting or analyzing scores so that only valid scores are used. It should ideally be done as soon as possible after the session and may involve discussion or review with the study team or PI. Decisions about reasons to consider a score invalid are up to the examiner/researcher; ideally, the examiner's or research study's general criteria should be in place prior to testing, but it is also necessary to use information from the test session when making these decisions, as each case is unique. Unexpected validity questions or test session challenges may come up. Additionally, if tests are marked as questionable validity (not valid or invalid), these cases and notes will need to be reviewed and a decision will need to be made. The Administration Form is a standardized form for documenting the most common issues that we have found, and it reflects the general criteria that we have decided on for determining validity. If needed, it can be adjusted to fit study-specific protocol.

The next step in evaluating scores is to check for the presence of scores at the floor or ceiling. In our experience testing with IDD we have found that List Sorting Working Memory and Picture Sequence Memory are sometimes impacted by this problem. In List Sorting Working Memory, the examinee sometimes passes practice but gets no sequences correct, thus getting a raw score of 0. We recommend using that score (considering it valid), as the examinee did demonstrate understanding of the task by passing practice, but this is a decision for the examiner/researcher to make. In Picture Sequence Memory, floor and ceiling scores can be harder to monitor for and to check, as the lowest and highest possible theta score depends on the test version. It is also difficult during testing to monitor if the examinee is getting everything correct or incorrect. Because there are multiple age versions, we recommend that if an examinee receives a ceiling score, they be given the next oldest age version (if available), and in the case of a floored score, they be given the next lowest age version (if available). For example, if an examinee has a score at the floor of Picture Sequence Memory Age 5-6, they should then receive Picture Sequence Memory Age 3-4. If an examinee receives a floored score on Picture Sequence Memory Age 3-4, there is no younger version available, so they would be done with Picture Sequence Memory. If an examinee has a score at the ceiling of Picture Sequence Memory Age 3-4, they should then receive Picture Sequence Memory Age 5-6.

Next, the examiner needs to select which test version score to use if the examinee received multiple versions of one test and thus has results for both versions in their report or data files. For instance, if the examinee failed practice on List Sorting Working Memory Age 7+ and was then given List Sorting Working Memory Age 3-6, the examiner needs to be sure to ignore that examinee's results from the List Sorting Working Memory Age 7+ and only use the scores from Age 3-6. For Picture Sequence Memory, if the examinee was given two age versions, the examiner needs to select which version's scores will be used. For example, if the examinee was

T-PAL
Translational Psychophysiology
and Assessment Laboratory

given Picture Sequence Memory 7+ and got a score at the floor, the examiner should have also given them the Age 5-6 version. After testing, the examiner should be sure to use the score from the Age 5-6 version and as always, be aware of what test version the score came from. It may be helpful to also keep a record of the examinee's additional score on the other version.

Once the valid scores have been identified for the examinee for each test, the examiner can determine if composite scores are possible for that examinee. The composites generated by the Toolbox software (available along with the test scores) will be present if all tests that comprise each composite score were completed (i.e., if practice was passed and the test was not skipped). The Fluid Composite will be present if there is a score from DCCS (normal test), Flanker (normal test), List Sorting Working Memory, Picture Sequence Memory, and Pattern Comparison Processing Speed. The Crystallized Composite will be present if there is a score from Picture Vocabulary and Oral Reading. The Cognitive Function Composite will be present if there is a score from each of these seven tests. The Early Childhood Composite will be present if there is a score from DCCS (normal test), Flanker (normal test), Picture Sequence Memory, and Picture Vocabulary.

There are a few things to keep in mind regarding the composite scores. If the examinee was given two versions of a test within one administration (such as Picture Sequence Memory 3-4 and Picture Sequence Memory 5-6), the composite is based on only one of these test versions; there is no way to select which version goes into the composite creation. This may be a problem, such as if the examinee got a ceiling score on Picture Sequence Memory 3-4. Floor and ceiling scores are another issue to keep in mind when interpreting the composites. As long as a score was received on the necessary tests, the composite will be generated. Additionally, since experimental versions do not have standard scores, the use of List Sorting Working Memory Age 3-6 with Extended Instructions will mean that the Fluid and Cognitive Function Composites are not generated. Finally, the validity of the examinee's test scores must be considered. If an examinee receives a score on every test, but one or more scores were considered invalid by the examiner, the affected composite score(s) may also be considered invalid. For instance, an examiner might give extra prompting during Flanker and decide that test was invalid (though a score was generated because the test was completed). The Fluid and Cognitive Function Composites would have then utilized that Flanker score and thus could be somewhat inflated.

Some researchers have used a different requirement for the number of tests making up each composite, instead requiring only 4 of 5 fluid tests and 1 of 2 crystallized tests for the composites.[8] This requirement would benefit individuals with IDD in that a composite would be available more frequently. We have used an alternative scoring method to calculate composites from deviation scores using these less strict criteria (4 of 5 Fluid and 1 of 2 Crystallized tests required).[9] Deviation based scores have been shown to perform very well for individuals at or near the floor of a cognitive assessment.[10] We utilized this method for Toolbox composites in our sample to avoid potential issues with the scores of low-performing examinees; other researchers may find this method useful. Like we have done with the SB5 as seen in Sansone et al. (2014) and in the SB5 software,[2] we create z-scores for each Toolbox test based on normative means and SDs for that examinee's chronological age band. We transform the z-scores into the composite metric. We utilize this method to create deviation-based Fluid, Crystalized, and Cognitive Function composites. Future study will evaluate the relative validity of deviation composites and the standard composites generated by the Toolbox software.
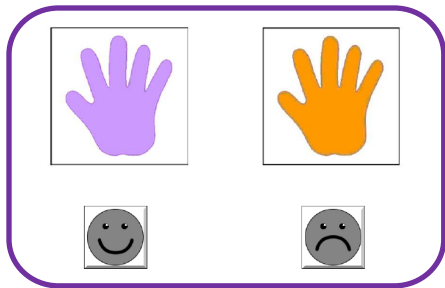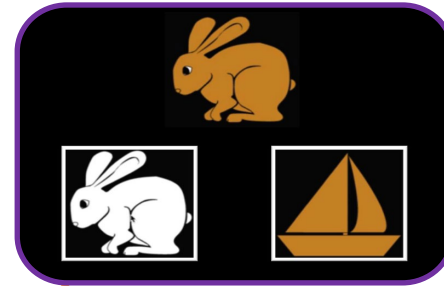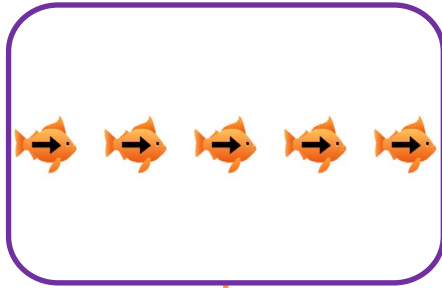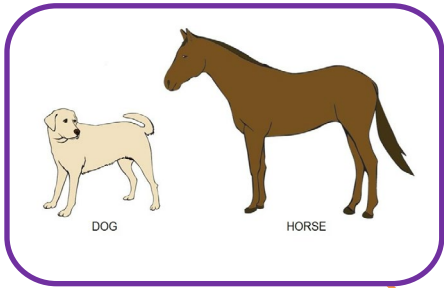
References

1.   National Institutes of Health and Northwestern University. NIH Toolbox for Assessment of Neurological and Behavioral Function Administrator's Manual: iPad App Version 1.17 and up. 2018.

2.   Roid G, Miller L. Stanford-Binet Intelligence Scales, Fifth ed. San Antonio: Pearson, 2003.

3.   Thompson T, Coleman JM, Riley K, et al. Standardized Assessment Accommodations for Individuals with Intellectual Disability. Contemp Sch Psychol 2018;22:443-457.

4.   Gray CA, Garand JD. Social stories: Improving responses of students with autism with accurate social information. Focus on Autistic Behavior 1993;8:1-10.

5.   NIH Toolbox [online]. Available at: http://www.healthmeasures.net/score-and-interpret/interpret-scores/nih-toolbox.

6.   Slotkin J, Nowinski C, Hays R, et al. NIH Toolbox Scoring and Interpretation Guide: National Institutes of Health and Northwestern University 2012.

7.   Casaletto KB, Umlauf A, Beaumont J. Demographically Corrected Normative Standards for the English Version of the NIH Toolbox Cognition Battery. 2015;21:378-391.

8.   Akshoomoff N, Beaumont JL, Bauer PJ, et al. VIII. NIH Toolbox Cognition Battery (CB): composite scores of crystallized, fluid, and overall cognition. Monogr Soc Res Child Dev 2013;78:119-132.

9.   Hessl D, Sansone SM, Berry-Kravis E, et al. The NIH Toolbox Cognitive Battery for intellectual disabilities: three preliminary studies and future directions. J Neurodev Disord 2016;8:35.

10.   Sansone SM, Schneider A, Bickel E, Berry-Kravis E, Prescott C, Hessl D. Improving IQ measurement in intellectual disabilities using true deviation from population norms. J Neurodev Disord 2014;6:16.
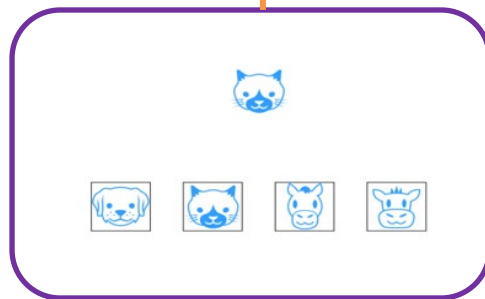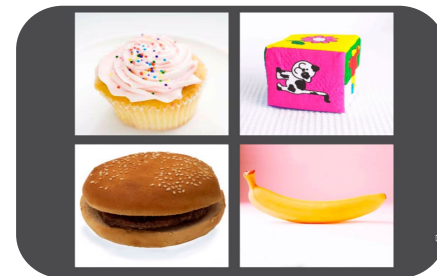
Reinforcement Inventory

Instructions: Below is a list of some items and experiences your son/daughter might find rewarding. Rate each according to how much your son/daughter "enjoys" the described item or activity and whether you will allow them to have it during their study visit. Also, it would be very helpful to get input from your son/daughter while completing the form.

Rewards	NOT AT ALL ☹	NOT REALLY ☹	NEUTRAL 😐	SOMEWHAT 🙂	VERY MUCH 🙂
Snacks					
1. Candy – List specific kinds below					
a.					
b.					
2. Chips – List specific kinds below					
a.					
b.					
3. Goldfish Crackers					
4. Graham Crackers					
5. Popcorn					
6. Pretzels					
7. Cookies – List specific kinds below					
a.					
b.					
8. Specific beverage or soda					
a.					
b.					
9. Other favorite snack items - list below					
a.					
b.					
Toys or Crafts					
10. Coloring/drawing					
11. Arts & crafts					
12. Play dough					
13. Stickers					
14. Bubbles					

Rewards	NOT AT ALL 🙁	NOT REALLY 🙁	NEUTRAL 😐	SOMEWHAT 🙂	VERY MUCH 😊
15. Stuffed Animals					
16. Dolls					
17. Cars					
18. Super heroes					
a.					
b.					
19. Other small toys / craft items – list below					
a.					
b.					
Entertainment/Activities					
20. Movies or TV programs– List below					
a.					
b.					
21. Music					
22. Books/Magazines					
23. Board games/puzzles					
24. Animals					
25. Sports – list below					
a.					
b.					
26. Other small activities – list below					
a.					
b.					
Items older participants might enjoy					
27. Nail polish/lip gloss					
28. Coffee gift card					
29. Shopping gift card					
30. Deck of cards/card games					
31. Comic books					

DOG   HORSE

Reading
7   5

D   3

Version 1.0
8/6/19

T-PAL
Translational Psychophysiology
and Assessment Laboratory

National Institutes of Health Toolbox Cognitive Battery Administration Form

Subject ID:	Date:
Examiner:	Visit #:

Order	Test	Passed practice?	Test valid?	Reasons invalid or unsure	Notes
	DCCS Was Home Base used? Y N N/A	Y N	Y N ? **DEXT**: Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____ _____ _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____ _____ _____ _____
	Flanker Arrows attempted? Y N Was Home Base used? Y N N/A	Y N	Y N ? **DEXT**: Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____ _____ _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____ _____
	Picture Vocabulary	N/A	Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____

Pattern Comparison	Y N	Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____ _____ _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____ _____ _____		
List Sorting Working Memory	Y N	Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____ _____ _____ _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____ _____ _____		
Oral Reading	N/A	Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____ _____ _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____ _____		

	Picture Sequence Memory Check the box for the first version given (if mental age is 3:0 to 6:11, administer Age 3-4 first and then the test portion only on Age 5-6): ☐ Age 3-4 ☐ Age 7 ☐ Age 8+	Y N	Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____ _____ _____ _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____ _____
N/A	**Picture Sequence Memory Age 5-6** (This section applies if participant has mental age between 3:0 and 6:11)	N/A	Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____ _____ _____ _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____ _____
	Speeded Matching	Y N	Y N ?	☐ Excessive prompting ☐ Examinee refused ☐ Invalid patterned response ☐ Technical difficulties ☐ Illness/emergency ☐ Other Please describe any selections above: _____ _____ _____	☐ Questionable understanding of task (describe below) Other notes: _____ _____ _____ _____

Testing situation	Low	Fair	Moderate	High
Rapport with participant	0	1	2	3
Participant comfort level	0	1	2	3
Issues during administration	**Not at all**	**Infrequent**	**Often**	**Persistent**
Excessively talkative	0	1	2	3
Easily distractible	0	1	2	3
Aggressive/oppositional	0	1	2	3
Escape behaviors	0	1	2	3

Reasons test may be invalid:

Excessive prompting: Prompting occurs during testing portion that significantly exceeds manual guidelines. Examples: redirecting examinee to the task repeatedly, referencing "paying attention" during the test, teaching the examinee during testing

Examinee refused: Refusal can be verbal or nonverbal. Examples: aggression toward examiner, attempting to leave the test room, verbally refusing to complete the test, randomly choosing answer options to end the test sooner

Invalid patterned response: If a examinee displays a systematic response style that reflects random or inappropriate response to what the task is asking, this can be a reason for an invalid score. For example, on Pattern Comparison Processing Speed a examinee may use a rigid alternating pattern (smiley face, frown face, smiley face, frown face, etc.), in which the examinee is not answering if the items are the same, but is instead selecting choices in a meaningless style. If prompting the examinee does not break them out of this response style, their Pattern Comparison Processing Speed score would be considered invalid. Another example of an invalid patterned response would be a examinee who does not like the frown face and thus only selects the smiley face.

Note: This category does not relate to the correctness of responses as long as the examinee is performing the task. (In Flanker, for example, a examinee always selecting the direction of the incongruent flanking fish and not the direction of the middle fish would not be a reason to consider the test invalid. Instead, it would indicate that the examinee had poor performance.)

Technical difficulties: Equipment fails and examiner is unable to resume task

Illness/emergency: Examinee or examiner is physically ill, or there is a power outage, earthquake, tornado, or other emergency that interferes with the testing session

Other notes:

Questionable understanding of task: While not an automatic reason to consider a score invalid (as long as the examinee did pass practice), it can be helpful to document when an examinee seems to not understand the task. When questionable understanding is noted, it is helpful to describe the reason this is suspected for later reference.

The notes section can also be used to document any additional information about test administration or examinee performance. For example, on Picture Vocabulary if the examinee was responding impulsively and the examiner moved the iPad farther away and instructed the examinee to slow down, this could be noted.

Rating scales:

Rapport with examinee

T-PAL
Translational Psychophysiology
and Assessment Laboratory

(0) Low : Uncomfortable with examiner, avoids eye contact, doesn't respond to examiner's requests or follow along with examiner on iPad

(1) Fair: Somewhat uncomfortable with examiner, makes some eye contact, responds to examiner occasionally and follows along on iPad occasionally

(2) Moderate: Moderately comfortable with examiner, cooperates most of the time and follows along on the iPad. Adequate sense of reciprocity between examiner and examinee

(3) High: Very comfortable with examiner, cooperates with examiner's requests and follows along on iPad. Strong sense of reciprocity between examiner and examinee

Examinee comfort level

(0) Low: Highly anxious, unable to sit still/appear calm, obvious aversion to the task/environment

(1) Fair: Anxious with some difficulty staying calm, some aversion to the task/environment

(2) Moderate: Appears generally at ease for most of the time, occasionally smiles and talks with examiner

(3) High: Appears at ease almost all of the time, smiles and talks with examiner

Excessively talkative

(0) Not at all: Might talk with examiner between tasks, but easily transitions to test and talking is not a problem

(1) Infrequent: Occasionally engages with examiner or self-talks during test, but is not disruptive to task

(2) Often: Talks often during the test, occasionally interferes with examinee's understanding of task

(3) Persistent: Talks excessively during testing, interferes with examiner's ability to administer test and limits examinee's understanding of task

Easily distractible

(0) Not at all: Focuses on task with no prompting from examiner

(1) Infrequent: Occasionally needs redirection to the task, but is attentive most of the time

(2) Often: Needs some prompting from examiner to complete each task, might be playing with/attending to other items in the room

(3) Persistent: Needs constant redirection to complete each task, is playing with/attending to other items in the room.

Aggressive/oppositional

(0) Not at all: Is cooperative and compliant during testing

(1) Infrequent: Has 1-2 instances of minor aggression/opposition (such as taking iPad from examiner, pushing examiner, etc.) but is cooperative and compliant for the most part

(2) Often: Has several aggressive/oppositional outbursts during the test (such as throwing things, hitting examiner, etc.) but behavior is manageable and testing can continue

(3) Persistent: Has multiple aggressive/oppositional outbursts that interfere with testing and examiner's/examinee's safety is at risk

Escape behaviors

(0) Not at all: Is cooperative and compliant during testing

(1) Infrequent: Tries to escape 1-2 times during testing (such as exiting the app, taking the iPad, getting out of chair), but behavior is easily and quickly managed and testing resumes

(2) Often: Tries to escape several times by getting out of chair or leaving room and behavior disrupts testing

(3) Persistent: Is constantly trying to get out of chair or leave room and behavior is very difficult to manage