Test-Retest Reliability and Practice Effects on the NIH Toolbox Cognition Battery in Duchenne Muscular Dystrophy

Aaron J Kaat[1] & Mathula Thangarajh[2]

[1]Department of Medical Social Sciences, Outcome and Measurement Science Division; Northwestern University Feinberg School of Medicine, [2]Department of Neurology, Division of Child Neurology; Virginia Commonwealth University

Methodological Issue Being Addressed

Does the NIH Toolbox Cognition Battery have appropriate evidence for use as a performance-based clinical outcome assessment in Duchenne Muscular Dystrophy (DMD)?

Introduction

- The **NIH Toolbox Cognition Battery (NIHTB-CB)** is a low-burden assessment battery measuring both fluid and crystallized cognitive (Gershon et al., 2013).
  - Fluid Cognition Tests:
    - Dimensional Change Card Sorting (DCCS)
    - Flanker Inhibitory Control and Attention (FICA)
    - Picture Sequence Memory (PSM)
    - List Sorting Working Memory (LSWM)
    - Pattern Comparison Processing Speed (PCPS)
  - Crystallized Cognition Tests:
    - Picture Vocabulary (PV)
    - Oral Reading and Recognition (ORR)
- There is strong validity evidence for the NIHTB-CB in the general population and it is increasingly being used in clinical samples (c.f. Fox et al., 2022).
- **Duchenne Muscular Dystrophy (DMD)** is a rare genetic disorder characterized by progressive skeletal and cardiac muscle weakness.
  - However, there is a growing recognition that DMD also affects cognitive functioning, notably in executive functioning areas (Thangarajh et al., 2019).
  - Cognitive monitoring is a recommended component of the recommended DMD Care Considerations (Birnkant et al., 2018).
- **Future clinical trials within DMD should also consider non-motor outcomes, including and especially cognitive outcomes.** This study evaluates some of the psychometric properties of the NIHTB-CB within DMD to determine if it may be appropriate for future clinical trials targeting these outcomes.

Methods

Participants
- 29 Boys
- Age Mean=10.4 SD=4.2, range=4.5-27 years old
- Race/Ethnicity:
  - Hispanic Ethnicity, Any Race: n=7, 24%
  - Non-Hispanic White: n=21, 72%
  - Asian: n=1, 3%
- Recruited for a larger psychometric readiness study, evaluating the NIHTB-CB and other patient- and observer-reported outcome measures annually.

Design
- The NIHTB-CB was administered at baseline and re-administered after 2-6 weeks.
- Raw Scores (i.e., normative input scores) for individual tests and uncorrected standard scores for composites, and age-corrected standard scores for both tests and composites were calculated .

Analysis
- Mixed effects model with random intercept for participant.
- ICC used to index test-retest reliability.
- Fixed effect for test occasion represented practice effects on the original score metric.
- The SMD [$(\mu_2 - \mu_1)/\sigma_1$] provided an effect size for the practice effect.

Northwestern Medicine®
Feinberg School of Medicine

Results

The primary results are summarized in **Table 1**.

Feasibility
- Data completion varied from 24-29 participants (83-100%) for individual tests.
- Composites require complete data on **all** component tests, reducing completion to 15 or 16 cases (51-55%).

Test-Retest Reliability
- Raw scores exhibited higher reliability than the age-adjusted scores (median ICC=0.85 vs 0.68).
- The reliability of the three composites was also higher than the reliability of the individual measures.
  - Raw scores: median 0.89 vs 0.75
  - Age-adjusted scores: median 0.88 vs 0.66

Practice Effects
- Practice effects were negligible-to-small and nonsignificant in all cases.
- Tests with scoring models incorporating both accuracy and response time (i.e., DCCS and FICA) had some of the smallest practice effect SMDs.
- Tests with adaptive administration (i.e., PV and ORR) also had very small practice effect SMDs.
- Two tests—one of which is a memory test—had a small SMD (i.e., SMD >= 0.20), though the effect was nonsignificant herein.

Table 1. Reliability of the NIHTB-CB

Test / Composite	Number of Completers	Raw Scores			Age-Adjusted
		ICC	Practice Effect*	SMD	ICC
Total Cognition Composite	15	0.89			0.88
Fluid Composite	15	0.83			0.76
Dimensional Change Card Sort	28	0.78	0.0007	< 0.001	0.52
Flanker Inhibitory Control	27	0.93	0.009	0.004	0.69
Picture Sequence Memory	29	0.73	18.16	0.15	0.58
List Sorting Working Memory	25	0.75	1.24	0.23	0.68
Pattern Comparison	25	0.75	3.13	0.20	0.66
Crystallized Composite	16	0.92			0.93
Picture Vocabulary	29	0.87	0.01	0.004	0.58
Oral Reading and Recognition	24	0.94	0.31	0.07	0.89

*Raw scores are not comparatively scaled across test; Unstandardized practice effects should not be compared. All practice effects p > 0.05

Conclusions

- Test-retest reliability on the NIHTB-CB was high and practice effects were nonsignificant, particularly for measures of **executive function.**
  - Subtle executive functioning deficits previously have been shown to be most relevant to individuals with DMD (Thangarajh et al., 2019).
- Feasibility for the composites was low.
  - This does not appear to be driven by any one test in particular.
  - Affecting cognitive domains globally (as opposed to specific cognitive skills) is also less likely so this may not be a major concern.
- Reliability was lower for norm-referenced scores, further supporting arguments against the context of use for norm-referenced scores as a clinical endpoint (c.f., Farmer et al., 2023).
- Nonetheless, this study provides initial support for the monitoring context of use for the NIHTB-CB—especially executive functioning tests—among those with DMD. More evidence is necessary, however, to further support this claim.

Limitations

- Missing data on the composite scores were missing not at random, so although the test-retest reliability was exceptional, it was only computable for individuals for whom the tests were feasible.
- The detection of practice effects is potentially underpowered (particularly for tests with small SMDs).

Future Directions

- Thorough evaluations of fit-for-purpose require more evidence than simply test-retest reliability and practice effects; gathering this evidence is an important next step.
  - Qualitative and mixed methods data are necessary.
  - The context of use anticipated here is monitoring (treatment response), but additional evidence may support other contexts of use as well.
- The NIHTB-CB version 3 was released in 2023, including new person-ability scores which may be more relevant for clinical trial use (c.f., Farmer et al., 2023). Future studies should evaluate the updated version.

References

Birnkrant, D. J., Bushby, K., Bann, C. M., Apkon, S. D., Blackwell, A., Colvin, M. K., … & DMD Care Considerations Working Group (2018). Diagnosis and management of Duchenne muscular dystrophy, part 3: primary care, emergency management, psychosocial care, and transitions of care across the lifespan. The Lancet. Neurology, 17(5), 445–455. https://doi.org/10.1016/S1474-4422(18)30026-7

Farmer, C., Thurm, A., Troy, J.D., and Kaat A.J. (2023), Comparing ability and norm-referenced scores as clinical trial outcomes for neurodevelopmental disabilities: a simulation study. *J Neurodevelop Disord* **15**: 4. https://doi.org/10.1186/s11689-022-09474-6

Fox, R. S., Zhang, M., Amagai, S., Bassard, A., Dworak, E. M., Han, Y. C., ... & Gershon, R. C. (2022). Uses of the NIH Toolbox® in clinical samples: a scoping review. Neurology: Clinical Practice, 12(4), 307-319. https://doi.org/10.1212/CPJ.0000000000200060

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH toolbox for assessment of neurological and behavioral function. Neurology, 80(11 Supplement 3), S2-S6. https://doi.org/10.1212/WNL.0b013e3182872e5f

Thangarajh, M., Kaat, A.J., Bibat, G., Mansour, J., Summerton, K., Gioia, A., Berger, C., Hardy, K.K., and Wagner, K.R. (2019), The NIH Toolbox for cognitive surveillance in Duchenne muscular dystrophy. *Ann Clin Transl Neurol*, **6**: 1696-1706. https://doi.org/10.1002/acn3.50867

Disclosures

The authors report no conflicts of interest for this work. Research reported in this poster was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number R21TR004007. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author CRediT Statement

AJK – Methodology; Formal Analysis; Data Curation; Writing—Original Draft, Review & Editing

MT – Conceptualization; Methodology; Investigation; Resources; Data Curation; Writing—Review & Editing; Funding Acquisition; Project Administration