Symposium Title: From Behavior to Biomarkers: Challenges and Advances in Outcome Measures for IDD

Co-Chair: Audrey Thurm¹ and Cristan Farmer¹

Discussant: Elizabeth Berry-Kravis

Overview: The selection of outcome measures is an important design consideration in any clinical trial. When the measured traits are at the extremes, such as IQ in samples of individuals with genetic syndromes associated with intellectual disability, investigators are faced with a slew of considerations beyond the basic psychometric characteristics of instruments. Is there one instrument that can be used for the age and ability range of the population? Can the instrument be used as intended in the population? Are alternative scoring methods appropriate? Does the test measure the construct the same way in this population as in the population in which it was developed? Alternatively, we have been encouraged by funding agencies to move towards the use of biomarkers as outcomes. Although the search for useful biomarkers in most intellectual and developmental disabilities (IDD) is ongoing, some have shown initial promise. The goal of this symposium is to encourage critical thinking among investigators as they select outcome measures, and to communicate exciting new data about biomarkers relevant to IDD. To accomplish this, we present viewpoints from six research groups on three distinct datasets/studies. The first talk uses acquired data from a large, longitudinal study of children with ASD to communicate the impact of one small decision—whether to use standard scores, age equivalents, or raw scores from the same instrument—on the change over time that should be expected during a study period. The second talk reports results from an essential, basic, yet often overlooked, psychometric analysis: the assessment of measurement invariance, and uses data from the Simons Simplex Collection to raise questions of whether popular ASD instruments work consistently across the broad spectrum of individuals, including those with IDD. Our third talk transitions to biomarkers as outcome measures, presenting new data supporting the use of a biomarker for a rare genetic disorder as an outcome measure, and also addresses the limitations of using standardized behavioral measures as an index of utility for biomarkers.

Paper 1 of 3

Paper Title: Outcome Measure Selection for Samples with ASD and Intellectual Impairment: A Vineland-II Vignette

Authors: Cristan Farmer, Ph.D.¹, Jesse Troy, Ph.D.² Audrey Thurm, Ph.D.¹

Introduction: The outcome measure is a key selection in any clinical trial with development as an outcome. In addition, multiple scoring options may be available, including: standard scores (SS), age equivalents (AE), or raw scores. Especially in samples with extreme scores on the outcome measure (i.e., very low cognitive or adaptive behavior scores), each score type has attendant advantages and limitations. For example, SS provide relative performance, but do not capture fine degrees of change, and may be censored at extreme values. AEs also provide some degree of relative performance, though wide intervals may prevent the capture of change (i.e., a scale may progress in 3-month chunks, so an improvement of 1-month may be missed). Raw scores promise the most potential sensitivity to change, but are difficult to interpret, especially across a wide age range, and may not be ratio-level. We employed the Vineland Adaptive Behavior Scales, Second Edition (VABS-II) Socialization subscale to illustrate the effects of score type selection and the influence of age and cognitive impairment, for consideration in study planning.

Method: Data are drawn from a longitudinal natural history study of 105 children with DSM-IV-TR Autistic Disorder (82% male), aged 2 to 7 years at baseline. Follow-up visits occurred at intervals of 6 months or 1 year (total 357 assessments). This autism cohort was enriched with children with intellectual impairment, with low baseline mean scores in VABS-II Adaptive Behavior Composite (65.70±9.98) and nonverbal developmental quotient (59.11±17.68). We characterize the short-term natural history of

¹ National Institute of Mental Health
² Duke University School of Medicine
VABS-II Socialization scores (SS, average domain AE, and raw total) as a function of child age and cognitive level. This was accomplished by using two related mixed models to (a) evaluate the shape of the average developmental trajectory across the study period (i.e., growth model with random intercept and slope) and (b) determine the average change in scores across a 6-to-18-month follow-up period (i.e., repeated measures model with random intercept).

**Results:** Results of the growth curve analyses demonstrated that VABS-II Socialization SS decreased significantly over 2-to-7 years of age, but this was moderated by cognitive level, such that larger decreases on SS were observed among individuals with lower cognitive scores. A significant quadratic effect was observed, suggesting that the expected rate of change in SS depended on the age of the child. Socialization AEs and raw scores increased significantly over the study period, but the slopes were moderated by cognitive level, such that individuals with less cognitive ability exhibited smaller increases. The linear model was best fit to the AE and raw-score data, suggesting an even rate of change across the age span. Results of the repeated measures analyses demonstrated that both age at assessment and cognitive level influenced the average amount of change observed in SS at approximately 1-year follow-up (e.g., at age 3 years, cognitive scores less than 50 were associated with a decrease of 5 standard units, but cognitive scores greater than 70 were associated with no change in standard score; conversely, at age 6, low cognitive scores were associated with no change in standard score, while high cognitive scores were associated with a significant increase of 3 standard units). Only cognitive level, not age, influenced the average amount of change observed in AE (e.g., no statistically significant change when cognitive scores were less than 50, and a significant increase of 7 months when cognitive scores were greater than 70) and raw scores (e.g., no statistically significant change when cognitive scores were less than 50, but a significant 12-point increase when cognitive scores were greater than 70).

**Discussion:** The shape of a function, or the developmental trajectory, in a given population is of intense interest when the course is likely to differ based on phenotypic characteristics such as cognitive impairment. In this presentation, we demonstrate that a variety of approaches exist to quantify a particular construct with one instrument, and that the expected trajectory of the construct may depend on how it is measured (e.g., SS, AE, raw score) and in whom it is measured (e.g., age and cognitive level of the participant). For example, investigators planning a study with a wide age range may consider that expected change over 1 year in SS may depend on the age of the child, as those planning a study in children with a wide range of cognitive ability should consider that the less cognitively able participants may naturally exhibit no change in AE or raw score over the study period. Thus, investigators are wise to consult existing data to explore which formulation best suits the intended purpose, an endeavor that is becoming more feasible in the era of data sharing (e.g., National Database for Autism Research).

**Paper Title:** The Meaning of SRS-2 Scores is Relative: The Importance of Measurement Invariance in Outcome Measure Selection

**Authors:** Vanessa Hus Bai3, Alexandra Havdahl4, Somer Bishop3, Audrey Thurm1, Cristan Farmer1

**Introduction:** Implicit in any comparison of scores between groups is the assumption that the instrument is measuring the construct in the same way across groups. This need not be an assumption, and when empirically tested and confirmed, the instrument is said to have the property of measurement invariance. Despite being an essential psychometric property of an instrument, measurement invariance is often overlooked. There are several levels of invariance, including configural, metric (sometimes called weak), and scalar (sometimes called strong). Configural invariance demonstrates that the overall model structure is the same between groups, metric invariance demonstrates that the factor loadings are equivalent across groups, and scalar invariance demonstrates that the intercept of each item is the same across groups. Without scalar invariance, factor scores cannot be compared, because the score has a different meaning between groups—it reflects the level of the latent construct, but also the influence of unmeasured or relevant factors that differ between groups. The goal of this analysis was test the

---

3 University of California, San Francisco
4 Norwegian Institute of Public Health, Oslo, Norway
assumption of measurement invariance in the use of the Social Responsiveness Scale, Second Edition across children with ASD and varying levels of language and cognitive ability. Several items on the SRS-2 relate to verbal abilities (e.g., conversation or tone of voice), suggesting that scores for minimally verbal children may need to be interpreted with caution. Given that the treatment subscales of the SRS-2 are not empirically derived, we tested the measurement invariance of the five-factor solution proposed by Frazier et al. (2014), comparing children with nonverbal IQ less than 70 and no phrase speech to children with nonverbal IQ greater than 70 and phrase speech.

Methods: Data were drawn from the Simons Simplex Collection, a permanent repository of genetic and behavioral data from over 2,600 simplex families. The sample was divided into those with nonverbal IQ greater than 70 and phrase speech according to the ADI-R \( n=2,338 \), and those with nonverbal IQ less than 70 and less than phrase speech \( n=298 \). Invariance was tested in a multiple group confirmatory factor analytic framework, using MPlus. The five-factor model was constructed using the same procedures as Frazier et al. (2014; personal communication, 2016). Model fit, and change in model fit as a result of increasing constraints on parameters between groups, was assessed with several statistics that are complementary in their strengths/weaknesses related to model complexity and sample size (in the interest of space, only selected statistics are reported in this abstract). Standard rules of interpretation were used for these fit indices (i.e., Hu & Bentler, 1999; Meade et al., 2008).

Results: The unconstrained CFA had a fair fit to the data, very similar to the fit indices in the original report (Frazier et al., 2014) (e.g., RMSEA 95% CI .075-.081, compared to Frazier et al.’s .08-.083). Thus, assuming configural invariance, metric invariance was tested by constraining the factor loadings to equivalence across groups. This did not significantly worsen model fit (e.g., \( \Delta \text{McNCl} = -.002 \)), indicating support for metric invariance. We proceeded to test scalar invariance, by constraining indicator intercepts to equivalence across groups, and found that this did significantly worsen model fit (e.g., \( \Delta \text{McNCl} = .048 \)). In other words, scalar invariance between subgroups with and without IQ>70 and phrase speech was not supported.

Discussion: Measurement invariance is assumed in all analyses of a measure where two groups are compared or combined; yet it is often overlooked. This becomes especially tenuous in situations where invariance might reasonably be questioned; as in the case of the SRS-2 across subgroups with and without IQ>70 and phrase speech. The present analyses demonstrated that invariance was not supported across these subgroups. Functionally, this means that it would be inappropriate to use the five-factor Frazier solution in the SSC sample without accounting for subgroup membership, and that factor scores should not be compared across groups. More generally, however, we wish to use the example of the SRS-2 to encourage researchers to empirically consider the psychometric properties of outcome measures, as we note that this lack of invariance is likely true for any number of instruments commonly used in the ASD literature.

References/Citations:
Introduction: Smith-Lemli-Opitz syndrome (SLOS) is an autosomal recessive inborn error of cholesterol metabolism syndrome that has neurocognitive manifestations. The cholesterol precursors 7- and 8-dehydrocholesterol (7- and 8-DHC) are known biomarkers of this disorder. In this paper, we describe the discovery that these biomarkers correlate cross-sectionally with the degree of functional impairment observed in the individual, and present new data exploring the possibility that change in these biomarkers may be associated with change in functional impairment, thereby qualifying as important candidates for use in clinical trials.

Methods: Participants were drawn from two studies, a longitudinal natural history study of SLOS (NCT00001721) and a clinical trial of simvastatin in individuals with SLOS (NCT00064792). SLOS diagnosis was confirmed biochemically and molecularly, and the 36 participants ranged in age from 4 to 33 years. Ten participants provided two samples at least 1 year apart. Serum and CSF samples were collected within 2 weeks of the behavioral assessment, and all samples were analyzed for levels of cholesterol, 7-DHC, and 8-DHC. The neuropsychological evaluation included assessment of cognitive function (Stanford Binet or Mullen Scales of Early Learning full scale IQ, in ratio IQ form (MA/CA) if standardized scores not available) and adaptive behavior (Vineland Adaptive Behavior Scales, Second Edition Adaptive Behavior Composite). In the full sample, partial correlations among the biomarkers and behavioral ratings were calculated, controlling for age. To evaluate the within-subject relationship between the biomarkers and behavioral ratings over time, individual regression lines will be calculated for each longitudinal participant’s set of data, and raw data plots will be presented.

Results: In the cross-sectional analyses, serum cholesterol was moderately related to adaptive behavior ($r=.53, p<.01$), but not to IQ. 7-DHC was related to IQ ($r=-.48, p<.05$) but not adaptive behavior. 8-DHC was weakly, but not significantly, related to both. However, the commonly-calculated ratio of (7-DHC + 8-DHC) to cholesterol was related to both adaptive behavior ($r=-.50, p<.01$) and IQ ($r=-.62, p<.01$). In the CSF, nearly-zero correlations were observed between cholesterol and both adaptive behavior and IQ. However, 7-DHC correlated moderately with adaptive behavior ($r=-.58, p<.01$) and IQ ($r=-.51, p<.05$), as did 8-DHC, though only adaptive behavior was significant (adaptive: $r=-.46, p<.01$; IQ: $r=-.38, p>.05$). Again, however, the ratio of 7- and 8-DHC to cholesterol was robustly related to adaptive behavior and IQ (adaptive: $r=-.57, p<.01$, IQ: $r=-.56, p<.01$). Results of the longitudinal analyses are forthcoming.

Discussion: Biomarkers represent an exciting complement to behavioral outcomes in clinical trials, especially in well-defined disease cohorts. In the case of SLOS, both cholesterol and cholesterol precursor levels reflect the known disease process, which is an advantage over less well-defined disorders and syndromes. Still, the possibility exists that biomarkers which indicate disease state may not relate meaningfully with change in functional outcome, potentially because of limitations on psychometric properties of behavioral measures given the wide age range and functioning level observed. Further, our ability to demonstrate a relationship between change in biomarker and change in behavior depends on specificity of measurement in both domains, which is an omnipresent challenge in research concerning individuals with limited cognitive and adaptive skills.

References/Citations:

5 Kennedy Krieger Institute
6 National Institute on Child Health and Human Development