Running title: Ability score as outcome measure

Person ability scores as an alternative to norm-referenced scores as outcome measures in studies of neurodevelopmental disorders

Cristan A. Farmer, Aaron Kaat, Audrey Thurm, Irina Anselm, Natacha Akshoomoff, Amanda Bennett, Leandra Berry, Aleksandra Bruchey, Bruce A. Barshop, Elizabeth Berry-Kravis, Simona Bianconi, Kim M. Cecil, Robert J. Davis, Can Ficicioglu, Forbes D. Porter, Allison Wainer, Robin P. Goin-Kochel, Caroline Leonczyk, Whitney Guthrie, Dwight Koeberl, Jamie Love-Nichols, Eva Mamak, Saadet Mercimek-Andrews, Rebecca P. Thomas, Gail A. Spiridigliozzi, Nancy Sullivan, Vernon R. Sutton, Manisha D. Udhnani, Susan E. Waisbren, Judith S. Miller

Correspondence: Cristan Farmer, PhD
Email address: Cristan.Farmer@NIH.gov
National Institute of Mental Health
Neurodevelopmental and Behavioral Phenotyping Service
10 Center Drive
Bethesda, Maryland, 20892 USA

Author information:

Aaron Kaat, Ph.D., Northwestern University, IL, USA (aaron.kaat@northwestern,edu)
Audrey Thurm, Ph.D., National Institute of Mental Health, MD, USA (athurm@mail.nih.gov)
Irina Anselm, MD, Boston Children's Hospital and Harvard University, MA, USA (irina.anselm@childrens.harvard.edu)
Natacha Akshoomoff, Ph.D., University of California San Diego, CA, USA (nakshoomoff@ucsd.edu)
Amanda Bennett, M.D., MPH, Children's Hospital of Philadelphia, PA, USA (bennettam@email.chop.edu)
Leandra Berry, Ph.D., Baylor College of Medicine, TX, USA (lnberry@texaschildrens.org)
Aleksandra Bruchey, Ph.D., Lumos Pharma, Inc., TX, USA (abruchey@lumos-pharma.com)
Bruce A. Barshop, M.D., Ph.D., University of California San Diego, CA, USA (bbarshop@ucsd.edu)
Elizabeth Berry-Kravis, M.D., Ph.D., Rush University, IL, USA (elizabeth_berry-kravis@rush.edu)
Simona Bianconi, M.D., *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, MD, USA (simona.bianconi@nih.gov)
Kim M. Cecil, Ph.D., University of Cincinnati, OH, USA (kim.cecil@cchmc.org)
Robert J. Davis, Pharm.D., Lumos Pharma, Inc., TX, USA (rdavis@lumos-pharma.com)
Can Ficicioglu, M.D., Ph.D., Children's Hospital of Philadelphia and University of Pennsylvania, PA, USA (ficicioglu@email.chop.edu)
Forbes D. Porter, Ph.D., M.D., *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, MD, USA (fdporter@mail.nih.gov)
Allison Wainer, Ph.D., Rush University, IL, USA (allison_wainer@rush.edu)

Robin P. Goin-Kochel, Ph.D., Baylor College of Medicine, TX, USA (kochel@bcm.edu)
Caroline Leonczyk, Ph.D., Rush University, IL, USA (caroline_leonczyk@rush.edu)
Whitney Guthrie, Ph.D., Children's Hospital of Philadelphia, PA, USA (guthriew@email.chop.edu)
Dwight Koeberl, M.D., Ph.D., Duke University, NC, USA (dwight.koeberl@duke.edu)
Jamie Love-Nichols, M.S., M.P.H., Boston Children's Hospital and Harvard University, MA, USA (jamie.love-nichols@childrens.harvard.edu)
Eva Mamak, Ph.D., Hospital for Sick Children, ON, CA (eva.mamak@sickkids.ca)
Saadet Mercimek-Andrews, M.D., Ph.D., Division of Clinical and Metabolic Genetics, Department of Pediatrics, University of Toronto, The Hospital for Sick Children, ON, CA (saadet.andrews@sickkids.ca)
Rebecca P. Thomas, M.A., Children's Hospital of Philadelphia, PA, USA (rebecca.p.thomas@uconn.edu)
Gail A. Spiridigliozzi, Ph.D., Duke University, NC, USA (gail.spiridigliozzi@duke.edu)
Nancy Sullivan, Ph.D., Boston Children's Hospital and Harvard Medical School, MA, USA (nancy.sullivan@childrens.harvard.edu)
Vernon R. Sutton, M.D., Baylor College of Medicine, TX, USA (vrsutton@texaschildrens.org)
Manisha D. Udhnani, M.S., Children's Hospital of Philadelphia, PA, USA (udhnanim@email.chop.edu)
Susan E. Waisbren, Ph.D., Boston Children's Hospital and Harvard University, MA, USA (susan.waisbren@childrens.harvard.edu)
Judith S. Miller, Ph.D., University of Pennsylvania, PA, USA (millerj3@email.chop.edu)

Acknowledgments: A portion of this work was presented as a poster at the 2019 Annual Meeting of the International Society for Autism Research in Montreal, Canada and at the 2019 Autumn Conference of the International Society for CNS Clinical Trials in Copenhagen, Denmark. This work was supported by the Intramural Research Program of the National Institute of Mental Health (1ZICMH002961) and with funding from Lumos Pharma. We thank the Vigilan study families and study coordinators for making this research possible. ClinicalTrials.gov: NCT00298246, NCT02931682.

Disclosures: CAF, AK, AT, IA, NA, AB, AB, BB, RJD, FDP, AW, RPGK, CL, WG, DK JLN, EM, SMA, RPT, GAS, NS, VRS, MDU, and SEW have no disclosures to report. C. Ficicioglu, EBK, JM, SB, and KMC have served as advisory consultants to Lumos Pharma.

Abstract

Although norm-referenced scores are essential to the identification of disability, they possess several features which affect their sensitivity to ~~true~~ change. Norm-referenced scores often decrease over time among individuals with neurodevelopmental disorders who exhibit slower-than-average increases in ability. Further, the reliability of norm-referenced scores is lower at the tails of the distribution, resulting in floor effects and increased measurement error for individuals with neurodevelopmental disorders. In contrast, the person ability scores generated during the process of constructing a standardized test with item response theory are designed to assess change. We illustrate these limitations of norm-referenced scores, and relative advantages of ability scores, using data from studies of autism spectrum disorder and creatine transporter deficiency.

Key words: item response theory, ability score, outcome measures, neurodevelopmental disorder, floor effect

Introduction

Classical test theory (CTT), which presumes that an observed test score can be decomposed into a true score and measurement error, has been the traditional approach to test development, and persists as the most common approach within the intellectual and developmental disabilities. However, the normative scoring used in standardized neurodevelopmental testing is commonly based on item response theory (IRT), a mathematical model used to explain the relationship between latent constructs and their observable manifestations (see Cappelleri, Lundy, & Hays, 2014 for a comparison of CTT and IRT). In this approach, developers administer a test to a normative sample and then use IRT to estimate the difficulty of each item. This information is used to convert ~~raw scores~~response patterns into person ability scores which quantify the level of the construct for the examinee. Person ability scores are then transformed into norm-referenced scores (e.g., T-scores, scaled scores, standard scores~~), given the age of the individual and any other relevant factors.~~) within narrow age-bands. Age equivalents, commonly defined as the ability score corresponding to the median norm-referenced score (e.g., T-score = 50) within an age-group, may be also be derived from person ability scores.

An advantage of norm-referenced scores is that all test-takers are scored on the same scale. For example, an IQ of 100 has the same interpretation regardless of age, which is that the individual's performance is commensurate with the average performance of a large sample of their chronological-age-peers. However, IQ does not provide information about the absolute ability of the individual. Administered over time, IQ scores remain stable if an individual gains ability at the same rate as one's peers; norm-referenced scores change only if changes in ability dramatically exceed or fall short of age-based expectations. Among children with neurodevelopmental disorders, the development of ability is expected to lag behind age-based expectations. As a result, subtle change in ability is obscured by standard scores, which as a result may even decrease over time (Bishop, Farmer, & Thurm, 2015).

A second limitation of norm-referenced scores is the floor effect (Hessl et al., 2009). Most standardized tests provide norm-referenced scores up to four standard deviations below average (e.g., a standard score of 40). Too few members of a given age band score below this level to make possible the calculation of psychometrically sound norm-referenced scores. As a result, all individuals with extremely low ability relative to their chronological age peers receive the same norm-referenced score (the "floor" of the range of possible test scores), which obscures important variability. Even if an individual does obtain a score above the floor of the test, the reliability of norm-referenced scores decreases as they approach the floor. ~~This means that a small change in raw score, which would~~ Ability scores are not ~~affect an average~~limited by this, since they are calculated in the full normative sample. Despite these known significant limitations of norm-referenced ~~score, could dramatically shift a very low norm-referenced score.~~scores, they are commonly used as outcome measures in longitudinal study of neurodevelopmental disabilities. In this brief report, we use real-world data to illustrate for non-methodologists some of the limitations of norm-referenced scores and relative advantages of person ability scores.

Methods

For the purposes of illustration, and not to test any hypotheses relevant to the studies of origin, data were drawn from two separate studies. For Illustration 1, Vineland Adaptive Behavior Scale, Third Edition (Vineland-3) (Sparrow, Cicchetti, & Saulnier, 2016) data were drawn from an ongoing observational study of creatine transport deficiency syndrome (CTD), a rare, X-linked, metabolic condition associated with intellectual disability (ID) and other neurodevelopmental disorders (van de Kamp et al., 2013). This study was approved by the ethics committee at each site in the US and Canada. Twenty-nine participants in this study had available baseline and 6-month data. ~~The Vineland-3 produces person ability scores, called growth scale values (GSV), and norm-referenced (V-scale) scores for each subdomain.~~The Vineland-3 produces person ability scores derived from raw scores through the Rasch model, called growth scale values (GSV), and norm-referenced (V-scale) scores for each subdomain (see Sparrow et al., 2016 for a full description of standardization procedures). The Motor domain was excluded because norm-referenced scores are not available for participants older than 9 years.

For Illustration 2, Differential Ability Scales, Second Edition (Elliott, 2007a) data were drawn from a completed natural history study of participants with autism spectrum disorder (ASD) and non-ASD developmental delay which was approved by an NIH IRB (protocol 06-M-0102). Twenty-six participants who were administered the Early Years Battery of the Differential Ability Scale-II twice within a 1-year period were included. ~~Ability and T-scores (norm-referenced) on the Differential Ability Scale-II are available at the subdomain level.~~T-scores (norm-referenced) and Rasch model-derived ability scores based on raw scores from the Differential Ability Scale-II are available at the subdomain level (see Elliott, 2007b for a full description of standardization procedures).

Mixed models for repeated measures were specified in SAS/STAT 9.3. The average within-subject change was evaluated using a main effect of time (pre/post). Both the magnitude of an effect and its precision are relevant to statistical power; we present their ratio as a standardized mean difference ($Effect\ size\ d_z = \frac{Estimated\ Difference}{StandardError * \sqrt[2]{DegreesFreedom}}$). A varianceexplained type effect size (f^2) was also calculated and yielded the same interpretation (available from authors upon request).

Results

Illustration 1: Vineland-3

Figure 1 shows the floor effects which are often observed in samples with neurodevelopmental disorders. Floor effects were more common among the V-scale scores than the GSV. This is most evident on the Expressive subscale, where a huge range of ability (GSV on the X-axis) was assigned a V-scale score at the floor (on the Y-axis). In only one case (the Domestic subdomain) did the GSV exhibit a floor effect while the V-scale did not. The pre/post effect sizes in Vineland-3 GSV were more positive than those for the V-scale scores (Figure 2). While confidence intervals were wide, many V-scale scores decreased (negative effect sizes), indicating worsening relative to peers, while GSV indicated stability or improvement in ability over time.

Illustration 2: Differential Ability Scale-II

All Differential Ability Scale-II T-scores indicated stability or modest improvement relative to normative expectations, but the confidence intervals were generally centered at zero (Figure 2). In contrast, change was detected in each of the Differential Ability Scale-II ability scores, such that the 95% confidence intervals did not include zero.

Discussion

Standardized tests produce scores on several scales, among which norm-referenced scores are used most commonly in both clinical and research settings. Norm-referenced scores are useful for documenting criteria for diagnosis and/or treatment planning in the case of disability, but they obscure both between-individual differences and within-individual change in ability over time. Specifically, floor effects homogenize variability by assigning the same norm-referenced score to all ability levels below some threshold. In contrast, IRT-derived ability scores tend to preserve that variability and allow for differentiation at lower levels. Further, ability scores are more sensitive to change within an individual than are norm-referenced scores. Here we illustrated that because they measure change in ability rather than relative standing, the effect sizes for change in ability scores tend to be positive and larger than those for the norm-referenced scores. The small sample size in this report conferred wide confidence intervals; future work in larger samples should result in narrower confidence intervals and more marked differences between the scores. Larger effect sizes confer more power and require fewer participants. If the goal of a study is to document change within individuals or to evaluate correlates of that change, as in naturalistic or observational studies, ability scores may be from a statistical perspective an attractive alternative to norm-referenced scores. ~~Future directions of our work include simulation studies to provide formal statistical support for the use of ability scores in clinical trials.~~

~~Standardized tests produce other types of scores, including raw scores and age equivalents.~~A future direction of our work is a simulation study to provide formal statistical support for the use of ability scores in clinical trials. We plan to evaluate the conditions under which ability scores may be preferable to standard scores (or vice versa); in addition to assessing the impact of sample size, sample heterogeneity, length of follow-up, and other factors, we will be able to explicitly evaluate group differences which was not possible in the current dataset. However, standardized tests produce other types of scores, including raw scores and age equivalents, which were outside the scope of this short report. Each type of score has its own profile of strengths and weaknesses. Raw scores exhibit a great deal of variability, which is attractive from a statistical perspective, but they are not measured at an interval scale~~. Whereas raw scores are not inherently meaningful, age equivalents have an intuitive interpretation. However, age~~, so changes at different levels of the scale are not comparable. Age equivalents are also not measured at the interval level, and like standard scores, they exhibit decreased reliability at extreme values (Bracken, 1988). Thus, our future simulation work will also evaluate each of these types of scores in relation to ability scores, specifically for use as indices of change in neurodevelopmental disability.

Ultimately, the selection of an outcome measure for any trial should be driven first by theory, and whether the investigator is interested in measuring ability or ability relative to others. Even if the ability score is best-suited to the hypothesis, there are other factors to consider. First, ability scores do not necessarily increase linearly with age, which is one reason for the existence of norm-referenced scores. For example, changes in language ability would be expected to occur at a much higher rate among toddlers than among teenagers; the extent to which this is true in the population will depend on the type of developmental disorder. A related consideration is that the variability in ability scores is likely to increase over age (or ability level). This heteroscedasticity may violate certain statistical assumptions. Unlike normative scores, a standard error of measurement or other indicator of reliability is not often provided by the test publishers for person ability measures (and is not for the measures used here). Thus, when using ability scores, the age range and ability level of a sample must be carefully considered in the analysis and interpretation of any study. A second consideration is that ability scores are generally only available at the subscale level, whereas many investigators prefer to use composite constructs (e.g., verbal IQ). Third, given that an ability score is a unitless measure, it is specific to the subscale and test version, which may be consequential if a child ages out of a given test version during the trial; ability scores from two versions of the same test cannot be compared. Fourth, like raw scores, ability scores are not directly interpretable, and foundational work must be done to determine what type of change in a given ability metric should be considered clinically meaningful versus simply statistically significant. Finally, it is important to recognize that if floor effects in a test are caused not by norming procedures but by the lack of items at lower levels of ability (as likely occurred on the Domestic domain of the Vineland-3), then the test may be incapable of providing sufficient sensitivity to change regardless of which type of score is used.

While in this work we illustrate the use of both types of scores in the context of skills which are expected to grow over time, it is worth noting that ability scores are available for any measure which was derived through IRT~~.~~ (though we must also acknowledge that IRT is not yet commonplace for IDD-specific measures). As test developers continue to address the dearth of adequate measures for use in neurodevelopmental disorders with modern psychometric methods like IRT, person ability scores should be presented as a viable alternative or adjunctive to the norm-referenced scores. Depending on the purpose, population, and design of studies of neurodevelopmental disorders, person ability scores may provide greater statistical power as a study outcome because they are designed to measure change and because they are equally sensitive to differences at low and average levels of ability.

References

Bishop, S. L., Farmer, C., & Thurm, A. (2015). Measurement of nonverbal IQ in autism spectrum disorder: scores in young adulthood compared to early childhood. *Journal of autism and developmental disorders, 45*(4), 966-974.

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26*(2), 155-166.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648-662.

Elliott, C. D. (2007a). *Differential Ability Scales, Second Edition*. San Antonio, TX: The Psychological Corporation.

Elliott, C. D. (2007b). *Differential Ability Scales, Second Edition: Introductory and technical handbook.* San Antonio, TX.: Psychological Corporation.

Hessl, D., Nguyen, D., Green, C., Chavez, A., Tassone, F., Hagerman, R., . . . Hall, S. (2009). A solution to limitations of cognitive testing in children with intellectual disabilities: the case of fragile X syndrome. *Journal of Neurodevelopmental Disorders, 1*(1), 33-45. Retrieved from http://dx.doi.org/10.1007/s11689-008-9001-8. doi:10.1007/s11689-008-9001-8

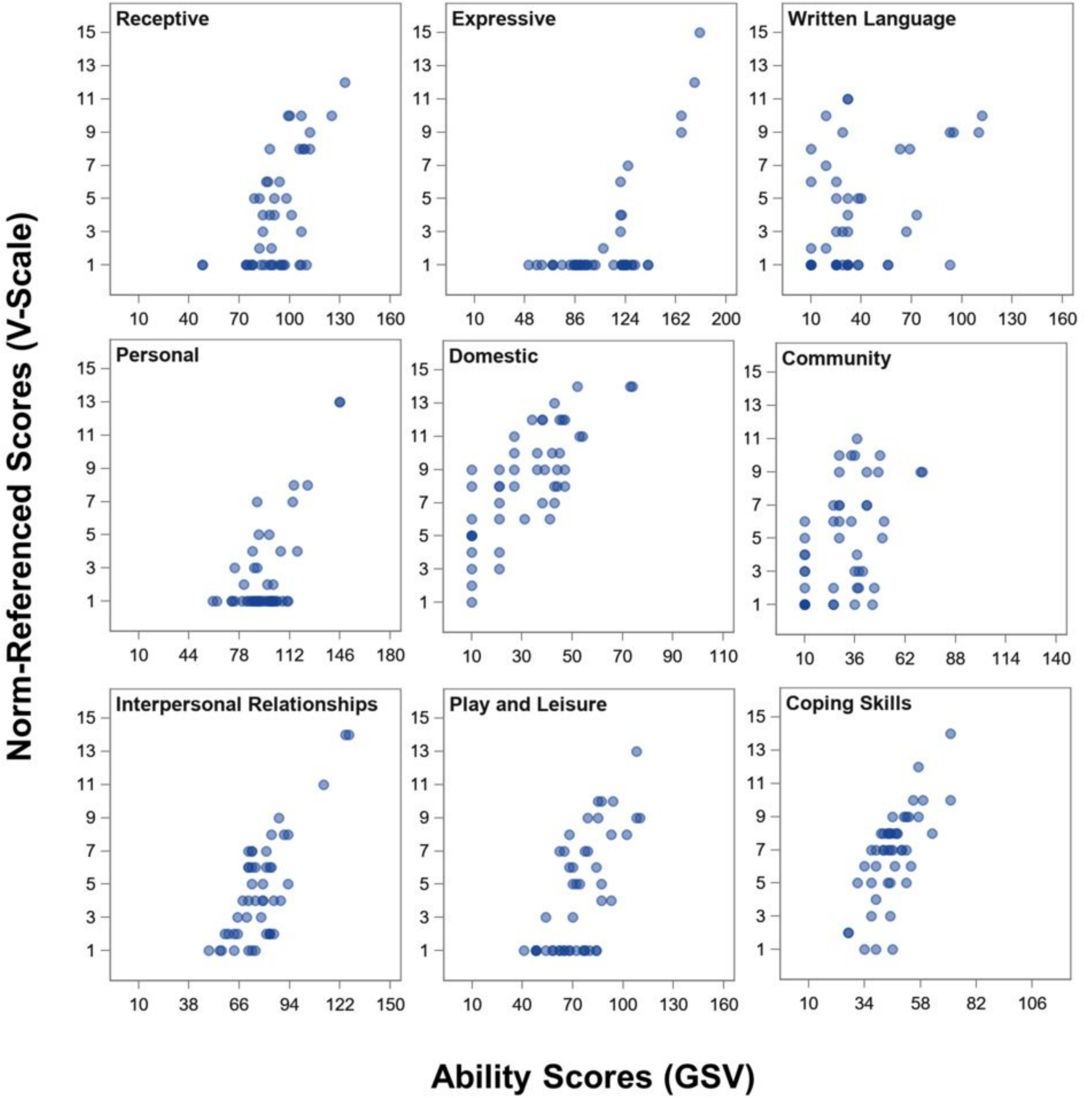
Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland Adaptive Behavior Scales, Third Edition*. San Antonio, TX: Pearson.

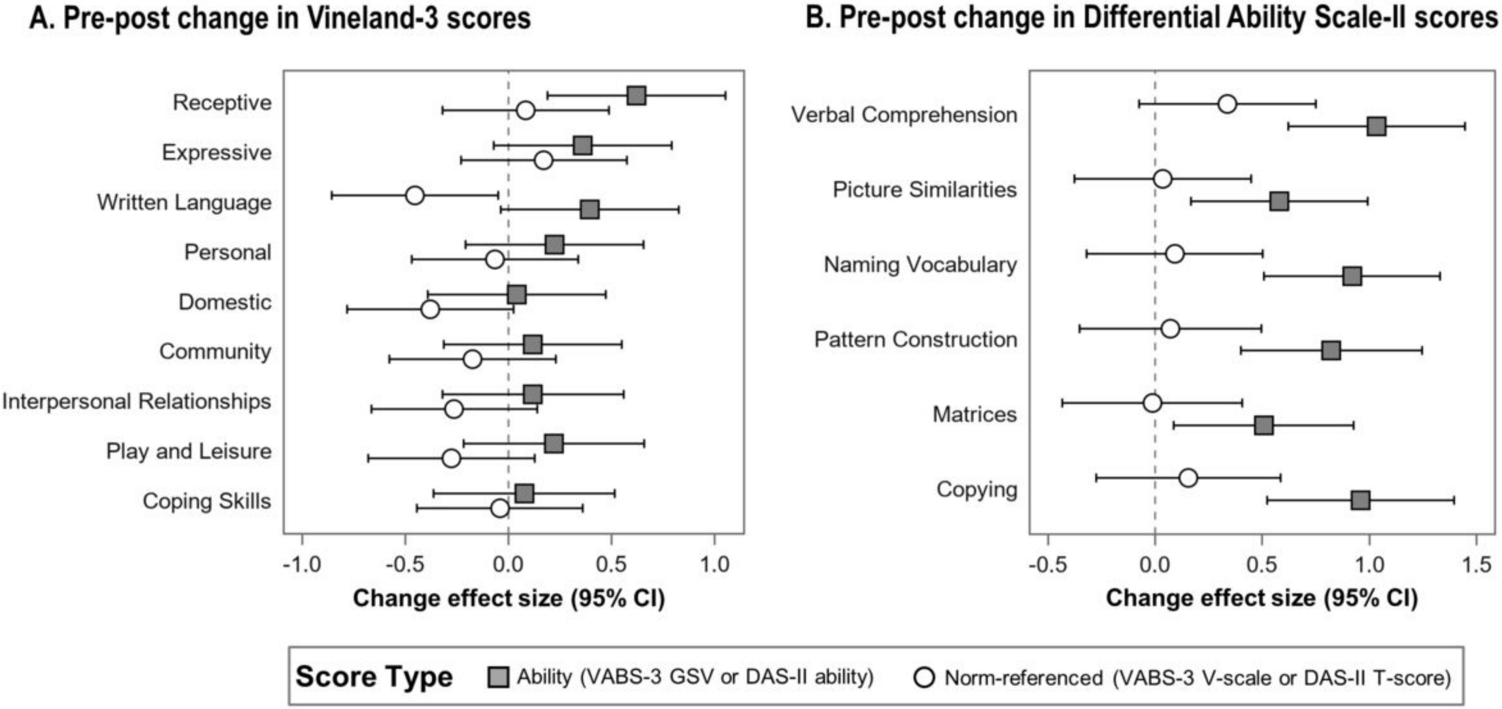
van de Kamp, J. M., Betsalel, O. T., Mercimek-Mahmutoglu, S., Abulhoul, L., Grünewald, S., Anselm, I., . . . Salomons, G. S. (2013). Phenotype and genotype in 101 males with X-linked creatine transporter deficiency. *J Med Genet, 50*(7), 463-472. Retrieved from https://jmg.bmj.com/content/jmedgenet/50/7/463.full.pdf. doi:10.1136/jmedgenet-2013-101658

Figure Legends

Figure 1. Floor effects in the Vineland-3. GSV = growth scale values. Baseline and 6-month Vineland-3 scores are plotted for N=29 individuals. Floor effects are characterized by "piling up" of markers at the lowest possible score on a scale. Markers are transparent; darker shades of blue indicate more observations at given location. Ranges of the X-axis reflect the actual range of possible GSV for a subscale.

Figure 2. Within-subject change effect sizes in ability scores versus norm-referenced scores. GSV = growth scale value; VABS-3 = Vineland Adaptive Behavior Scales, 3rd edition; DAS-II = Differential Ability Scale, Second Edition. Panel A shows the standardized effect size for baseline to 6-month change among 29 individuals. Panel B shows the standardized effect size for baseline to 1-year change among 26 individuals. For both panels, a negative effect size reflects a decrease in score and a positive effect size reflects an increase.


Norm-Referenced Scores (V-Scale)
Ability Scores (GSV)
Receptive
Expressive
Written Language
Personal
Domestic
Community
Interpersonal Relationships
Play and Leisure
Coping Skills
1
3
5
7
9
11
13
15
10
40
70
100
130
160
48
86
124
162
200
44
78
112
146
180
30
50
90
110
36
62
88
114
140
38
66
94
122
150
34
58
82
106


A. Pre-post change in Vineland-3 scores
Receptive
Expressive
Written Language
Personal
Domestic
Community
Interpersonal Relationships
Play and Leisure
Coping Skills
-1.0
-0.5
0.0
0.5
1.0
Change effect size (95% CI)
B. Pre-post change in Differential Ability Scale-II scores
Verbal Comprehension
Picture Similarities
Naming Vocabulary
Pattern Construction
Matrices
Copying
-0.5
0.0
0.5
1.0
1.5
Change effect size (95% CI)
Score Type
Ability (VABS-3 GSV or DAS-II ability)
Norm-referenced (VABS-3 V-scale or DAS-II T-score)