American Journal on Intellectual and Developmental Disabilities

WHICH SCORE FOR WHAT? OPERATIONALIZING STANDARDIZED COGNITIVE TEST PERFORMANCE FOR THE ASSESSMENT OF CHANGE

--Manuscript Draft--

Manuscript Number:	AJIDD-D-24-00047R1
Article Type:	Invitation Only - Developmental Synaptopathies
Keywords:	Person ability scores; Item Response Theory; Rasch analysis; Psychometrics; Longitudinal data
Corresponding Author:	Cristan A. Farmer National Institute of Mental Health Bethesda, MD UNITED STATES
First Author:	Cristan A. Farmer
Order of Authors:	Cristan A. Farmer
	Audrey Thurm
	Tanvi Das
	E. Martina Bebin
	Jonathan A. Bernstein
	Elizabeth Berry-Kravis
	Joseph D. Buxbaum
	Charis Eng
	Thomas Frazier
	Antonio Y. Hardan
	Alexander Kolevzon
	Darcy A. Krueger
	Julian A. Martinez-Agosto
	Hope Northrup
	Craig M. Powell
	Latha Soorya
	Joyce Y. Wu
	Mustafa Sahin
Manuscript Region of Origin:	UNITED STATES
Abstract:	Developmental domains such as cognitive, language, and motor are key concepts of interest in longitudinal studies of intellectual and developmental disabilities (IDD). Normative scores (e.g., IQ) are often used to operationalize performance on standardized tests of these concepts, but it is the interval-distributed person-ability scores that are intended for the assessment of within-individual change. Here we illustrate the use and interpretation of several Stanford Binet, 5th Edition score types (IQ, extended IQ, Z-normalized raw score, developmental quotient, raw sum score, age equivalent, and ability score) using data from two longitudinal studies of rare genetic conditions associated with IDD. We found that while normality assumptions were tenuous for all score types, floor effects led to model unsuitability for longitudinal analysis of most types of norm-referenced scores, and that the validity of interpretation with respect to individual change was best for ability scores.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Manuscript clean
Click here to access/download;Manuscript;rdcrn sb5 draft R1 clean.docx

WHICH SCORE FOR WHAT? OPERATIONALIZING STANDARDIZED COGNITIVE TEST PERFORMANCE FOR THE ASSESSMENT OF CHANGE IN

Abstract

Developmental domains such as cognitive, language, and motor are key concepts of interest in longitudinal studies of intellectual and developmental disabilities (IDD). Normative scores (e.g., IQ) are often used to operationalize performance on standardized tests of these concepts, but it is the interval-distributed person-ability scores that are intended for the assessment of within-individual change. Here we illustrate the use and interpretation of several Stanford Binet, 5th Edition score types (IQ, extended IQ, Z-normalized raw score, developmental quotient, raw sum score, age equivalent, and ability score) using data from two longitudinal studies of rare genetic conditions associated with IDD. We found that while normality assumptions were tenuous for all score types, floor effects led to model unsuitability for longitudinal analysis of most types of norm-referenced scores, and that the validity of interpretation with respect to individual change was best for ability scores.

Keywords

Standard scores, Age equivalents, Person ability scores, Sum scores, Item Response Theory, Rasch analysis, Psychometrics, Longitudinal data, Rare genetic conditions, Change Sensitive Score, Stanford Binet

Introduction

Developmental concepts such as cognition, motor skills, social and emotional abilities, and adaptive behavior are central to research on intellectual and developmental disability (IDD). Developmental measures that may have robust psychometric profiles when used in the general population may be insufficient for those with IDD, especially when the goal is not to identify disability but to monitor change in longitudinal research. Standardized assessments often have limited validity for populations with IDD because of the test floor; tests which are appropriate for the chronological age of an individual with IDD may be too difficult for the individual's developmental level, especially in the case of moderate-to-profound IDD. A common solution in this situation is to administer an out-of-age-range test, which necessitates the use of scoring methods alternative to norm-referencing, such as developmental quotients (DQs; the ratio of age equivalent to chronological age) (Soorya et al., 2018). The use of DQs is generally viewed as necessary but suboptimal, as they allow for the estimation of an individual's performance but have significant limitations. Especially for individuals at the extremes of the distribution, the DQ is a poor approximation of the IQ, and the discrepancy is inconstant across age (Ostrolenk & Courchesne, 2023). The meaning of change in DQ can be unclear because the denominator of chronological age continues to increase even after the numerator of mental age plateaus, leading to artifactual declines in DQ over time (Bishop et al., 2015). Further, age equivalents and by extension DQ are subject to a second type of floor effect, which is the youngest age at which the test is normed.

If an individual with IDD has sufficient ability to perform the easiest items of an assessment intended for their chronological age (i.e., exceed the test floor), then a norm-referenced score is possible. Norm-referencing is a key feature of many standardized developmental tests, as it is used to assist the interpretation of performance by comparing it to that of same-age peers. At the individual level, this allows for validity in the diagnostic context, because disability is typically defined relative to expected functioning. At the group level, norm-

referencing is intended to facilitate the valid comparison of performance across individuals when the effect of development is considered a nuisance. Norm-referenced scores for developmental tests, whether they are standard scores, T-scores, or scaled scores, usually express performance as a function of a normal distribution, and so the units of norm-referenced scores are standard deviations (SD) and the values correspond directly to percentiles (see **Table 1**). A standard score of 100 reflects performance that is as good or better than 50% of the population, a standard score of 55 is as good or better than 0.013% of the population, and so on. Because directly estimating a score at the <1st percentile requires a prohibitively large sample per normative group, scores more than about 3 SD below average are extrapolations (i.e., actual standardization data in this range may not be present; see Timmerman et al. (2021) for a tutorial on one type of regression-based norming procedures). Even after borrowing statistical information from adjacent age groups, the precision of the extrapolated values is low. Thus, by convention, scores more than 3-to-4 SD below average are usually censored. This lowest standard score offered by the publisher is the third type of floor effect, referred to here as the standard score floor.

Floor Effects: Challenges and Proposed Solutions

The test floor and standard score floor effects are of significant consequence in IDD because they have important consequences for the statistical analysis and interpretation of group-level effects (Wang et al., 2008). These effects are applicable to both the cross-sectional and longitudinal contexts. Test floors are important because if a test cannot be used for some proportion of a sample with IDD, the data are systematically missing and the results from the available testing will be biased positively. For those who can get past the test floor, the standard score floor obscures variability for only low scores. This reduces responsiveness to change in ability that occurs below the standard score floor and induces heterogeneity in variance (heteroscedasticity) that biases the estimated standard errors. This is diagnosed by a “conical” pattern in model residuals, where the variance in residuals increases as a function of the

predicted values. Similar to the effect of the test floor, standard score floors result in positively biased estimates of the intercepts and slopes with inaccurate standard errors, depending on the score region. As standard score floor effects increase as a function of age, they could cause artifactual nonlinearity in both the fixed and random effects (Wang et al., 2008) (though we note that the test floor effects may lessen as a function of age).

One proposed solution to standard score floor effects is the Z-score method (Hessl et al., 2009), wherein an individual's raw score is expressed as a function of the norm-group raw score mean and SD, with no censoring (see **Table 1**). While this method does remove the normative floor effects, major test publishers do not use it for at least two reasons. First, it rests on the assumption that raw scores are normally distributed within each normative age group. Skewness, which is often observed especially at the youngest and oldest ages, compromises this. When test developers do base standard scores on raw scores, this skewness is addressed by first normalizing, or converting to percentiles, the raw scores. Second, the Z-score method creates discontinuity at age breaks, such that one could observe a dramatic difference in Z-score for the similar performance across two adjacent age groups. For standard scores based on raw scores this is addressed with the statistical procedure of smoothing growth curves. Especially in IDD research, however, the benefit of estimating normative scores below the standard score floor may exceed the risk of these limitations.

Understanding Change

In the longitudinal context, the goal is to understand within-person change in the outcome of interest. Regardless of the method, change normative scores is challenging to interpret because it quantifies not only within-individual change in performance, but age-related differences in the normative sample. As a result, change in norm-referenced score has an indeterminate relationship with change in absolute levels of the underlying construct. A decrease in norm-referenced score can – though not necessarily – occur due to an actual decrease in skills (e.g., degeneration observed in many rare genetic conditions associated with

IDD). However, because skills in developmental constructs are expected to increase over time, decreasing normative scores can also occur when the acquisition of skills is slower than expected, or if skills are simply maintained. Thus, degeneration cannot be distinguished from slow gains or stability – a serious threat to the validity of the interpretation of change in norm-referenced scores. Thus, both floor effects and indeterminacy of change seriously threaten the validity of interpreting norm-referenced scores from a developmental test in the longitudinal context.

Many developmental tests do contain a scoring method intended for monitoring change (Farmer et al., 2022). These scores, called person-ability scores, are derived via item response theory or Rasch analysis (see **Table 1**). These approaches transform the ordinal raw sum score (or sometimes the pattern of item scores) into an interval-level measurement representing the ability that would produce that performance. Interval-level measurement means that a given difference in ability score has the same meaning at all points in the scale – this property is an essential assumption for most statistical models common to longitudinal data analysis and is required for the valid interpretation of the magnitude of resulting parameters. Because all Rasch-based and most IRT-based ability scores have a monotonic relationship with the raw sum score (the raw sum score is an ordered approximation of the ability score; Sijtsma et al., 2024), valid interpretation of the direction of change is possible. Finally, ability scores are subject only to the single test floor, and therefore have limited risk of flooring-related bias in parameter estimates described above. For these reasons, ability scores have the reverse functional profile of norm-referenced scores, well-suited to longitudinal but not diagnostic contexts.

Current Study

The range of options for expressing performance on any test (**Table 1**) presents an opportunity for researchers to select the option that is best fit for their intended purpose. We propose that the key elements to consider are the study population, the study design, and the

intended interpretation of the scores. Here, we focus on the case of longitudinal research in rare genetic conditions associated with neurodevelopmental disorder (GCAND), which are often associated with moderate-to-profound IDD. In contrast to cross-sectional studies, which focus on between-person differences, the goal of longitudinal research is to describe within-person change. Longitudinal research may be interventional, as in clinical trials, or observational, as in natural history studies. We use a common longitudinal analytic method, hierarchical linear modeling, to evaluate the relative profiles of each score type with respect to statistical and practical interpretation, and offer an appraisal of the validity argument for the use of each as an outcome in longitudinal research. While the principles discussed here should apply to any norm-referenced test of a developmental construct, here we focus on cognitive ability and the Stanford Binet, 5th Edition. We hypothesized that consistent with the background described above, we would observe statistical and theoretical limitations to the validity of model results using norm-referenced scores, and that the person ability score would have the most favorable profile of results. This report is intended to support clinical trial readiness by contributing to the literature base supporting the selection of person ability scores as endpoints for studies of individuals with IDD.

Methods

Participants

The Developmental Synaptopathies Consortium (DSC), a Rare Disease Clinical Research Consortium (https://www.rarediseasesnetwork.org/), comprises researchers conducting three multisite natural history studies of GCANDs: Phelan McDermid Syndrome (NCT02461420; see Levy et al., 2022), tuberous sclerosis complex (TSC) (NCT02461459), and PTEN hamartoma tumor syndrome (PTEN) (NCT02461446; see Busch et al., 2023; Busch et al., 2019). Phelan McDermid Syndrome is caused by a terminal 22q13.3 deletion encompassing the *SHANK3* gene or a pathogenic sequence variant in *SHANK3*, both resulting in haploinsufficiency. TSC is an autosomal dominant condition caused by loss-of-function

mutations in *TSC1* or *TSC2*. PHTS is a genetic condition caused by germline mutations in *PTEN*, which encodes phosphatase and tensin homolog. While the clinical manifestations of each of these conditions is heterogeneous, each is associated with IDD (among numerous other features). Because it is the score types and not the conditions which are the focus of this manuscript, we do not further describe the conditions themselves. Each study was approved by a centralized IRB, and informed consent was obtained from legal guardians, as well as assent where possible. The dataset was issued in October 2021. Participants in the dataset were included in the analysis if they had at least one assessment with the Stanford Binet, 5th edition.

Measures

The Stanford Binet, 5th edition (SB5) was refined using Rasch analysis and normed on a nationally representative sample of N=4,800 aged 2 to 85 years (Roid, 2003). There are 10 subtests which feed into the full-scale (FS) composite used in this study. The available score types for the SB5 FS are described in detail in **Table 1**; in the current study we evaluated the IQ, extended IQ (EXIQ), developmental quotient (DQ), Z-score (Z), raw sum score (RAW), age equivalent (AE), and change sensitive score (CSS). IQ, EXIQ, DQ, and Z are normative scores. RAW, AE, and CSS are absolute scores. CSS is the person ability score on the SB5; test publishers commonly apply a trade name to these scores (e.g., growth scale values on the Vineland Adaptive Behavior Scales).

The SB5 scoring process automatically generates the CSS, AE, and IQ scores. When an IQ is at the floor (40) or ceiling (160), the user may also choose to access EXIQ scores via the manual. DQ is calculated as AE*100 divided by the chronological age (in months). Z is calculated using the raw score and published age-group-specific means and standard deviations (Sansone et al., 2014). The script for score derivation can be obtained at [SEE SUPPLEMENTARY MATERIALS PROVIDED FOR REVIEW].

Statistical Analyses

To model within-person change in SB5 scores, we used hierarchical linear modeling. An identical but separate analysis was performed for each score type within each study. To account for clustering within participant, a random subject-level intercept (ID) was included in the model. A subject-level slope for DURATION (described below) was also included, reflecting variability across participants in their rate of change.

Age at baseline was highly variable and so the chronological age variable contained both between-subject (i.e., differences between older and younger participants) and within-subject (i.e., change within a participant over time) information. To differentiate between developmental and cohort effects, thereby avoiding the inferential error of attributing between-subject differences to the within-subject effect, chronological age was decomposed into two fixed effects (Curran & Bauer, 2011): time-invariant $\overline{AGE}$ (the participant's average age during participation, centered at the sample mean age of 11 years) and time-varying DURATION (the passage of time within a person, centered at the person's mean age; the slope for this term is referred to as "annualized change"). This disaggregation also creates more accurate (larger) estimates of variability in the fixed effects. A quadratic form was specified for $\overline{AGE}$, and in parallel, for DURATION via the DURATION*$\overline{AGE}$ interaction (i.e., within-subject change was allowed to depend on the participant's age). To allow for comparability across results, the same fixed ($\overline{AGE}$, $\overline{AGE}^2$, DURATION, $\overline{AGE}$*DURATION) and random effects (ID, DURATION) were specified for all score types. The within-subject terms (DURATION and $\overline{AGE}$*DURATION) correspond to the research questions of longitudinal research, specifically how much change was observed at the individual level. To aid interpretation of these within-subject parameters values, specific contrasts were used to estimate the fixed effect of DURATION for hypothetical participants at a representative range of ages (3, 7, 11, 15, and 19 years). R version 4.2.2 was used to implement the package lme4 (Bates et al., 2014). The R script can be obtained from [SEE SUPPLEMENTARY MATERIALS PROVIDED FOR REVIEW].

Model Fitting

Because they reflect the study design, all fixed and random effects were retained regardless of their contribution to the model. Both level 1 and level 2 model residuals were visually inspected for consistency with the required assumptions of normality and constancy. Across most models, the residuals departed from these assumptions in two important ways. First, a conical shape in residuals can be induced by floor effects, and second, an excess of extreme residuals can occur in the presence of unmodeled causal variables. Addressing non-normality is beyond the scope of this paper, but the identification of score types more likely to exhibit violations of modeling assumptions is of high relevance to the goal of comparing optimal scoring rules for a longitudinal context of use. For brevity, only our conclusions from visual inspection are included here, but residual plots can be found at [SEE SUPPLEMENTARY MATERIALS PROVIDED FOR REVIEW].

Statistical Interpretation

The magnitude and precision of fixed effects is described using the parameter estimates with 95% confidence intervals, as well as the associated test statistics and uncorrected p-values. To facilitate understanding of the results, here we review the statistical meaning of these parameters. The **intercept** for each model is the estimated point-in-time score for a participant whose average age of participation was 11 years ($\overline{AGE}$ = 0; this variable is grand-mean centered), at the middle of their study participation (DURATION = 0). The **estimate of $\overline{AGE}$** is interpreted as the expected between-subject *difference* in the outcome for an individual whose average age of participation is 1 year older than the group average; the **quadratic term for $\overline{AGE}$** allows for this between-subject difference to become smaller or larger for participants older or younger than the average. A negative slope for the quadratic term indicates that differences as a function of $\overline{AGE}$ are smaller (or more negative) for older participants than younger participants. If the main effect of $\overline{AGE}$ or its quadratic term is nonzero, then the estimated

average value (i.e., the intercept) depends on the average age of the participant. The **slope of DURATION** is interpreted as the within-person expected *change* in the outcome for each year of participation in the study, also known as the annualized change. An **interaction between $\overline{AGE}$ and DURATION** allows for the annualized change (DURATION) to depend on the person's mean age ($\overline{AGE}$); for example, older participants might gain skills more slowly than younger participants (a negative slope for the $\overline{AGE}$*DURATION interaction). If the linear and quadratic effects are similar between DURATION and $\overline{AGE}$, then one might interpret the estimated annualized change as applying for the full age range in the study. When DURATION and $\overline{AGE}$ are dissimilar, however, it is referred to as a cohort effect. This means that the estimated differences between participants are more or less than would be expected as a function of the estimated within-subject change. Cohort effects are especially relevant for cross-sectional comparisons or any use of the between-subject terms in longitudinal data, as differences between ages cannot be solely attributed to the effect of development.

Results

The Phelan McDermid Syndrome cohort was excluded from analysis because too few participants in the dataset had sufficient ability to take the SB5, resulting in a too-small sample size for the proposed analyses (n=24 out of 101 participants, several with only one assessment). Most participants in the PTEN dataset (n=91 of 107) and the TSC dataset (n=81 of 106) received at least one SB5 and were included in the analysis (**Table 2**). About half of the individuals without SB5 were reported to not have sufficient ability to take the test. The median number of yearly SB5 assessments per person in both studies was 3 [IQR: 2, 3].

The first available assessment from each person was used to illustrate the relationships amongst scores. Amongst the norm-referenced scores, Z yielded the largest estimates, followed by IQ and EXIQ. The EXIQ had the largest standard deviations, though the standard deviations for all norm-referenced scores exceeded 15 (the value in the population) (**Table 2**). In both

groups, DQ yielded a lower estimate than the norm-referenced scores. The norm-referenced scores (FSIQ, EXIQ, Z) and DQ were all very strongly and positively correlated ($\rho > 0.93$) with one another, and more moderately with the absolute scores (RAW, AE, and CSS; for PTEN $\rho = 0.62 – 0.67$ and for TSC $\rho = 0.35 – 0.46$) (**Figure 1**). The nearly perfect rank-order correlation amongst the absolute scores (RAW, AE, and CSS) is expected as by definition they have a monotonic relationship.

Statistical Interpretation

The raw data subjected to hierarchical linear modeling are shown in **Figure 2.** Here, we offer a narrative summary of this modeling (see **Table 3** for parameter estimates and test statistics and **Table 4** for summary). For PTEN, the older and less uniformly impaired of the two samples, the classic standard scores (IQ and EXIQ) would be interpreted as stable within person regardless of age because slope point estimates did not differ from zero (**Figure 3**), but the floor effects suggested that the model results might be biased. Because EXIQ simply replaced the floor values in IQ with a new floor (i.e., almost no scores between 40 and 10 were observed), it did not successfully address the censoring in IQ. The norm-referencing methods which mitigate floor effects, DQ and Z, did both result in lower estimated scores across the age range and did decline within person. These effects were presumably obscured by censoring in IQ and EXIQ, and so these models support the observation that parameter estimates from IQ and EXIQ models were biased. The absolute scoring methods, RAW, AE, and CSS, all indicated growth that was more rapid for younger participants and leveled off for older participants, consistent with expectations for a developmental trajectory (**Figure 3**). The AE data, however, exhibited both floor and ceiling effects that indicated the possibility of bias. RAW and CSS had excess positive residuals that could threaten model validity. Most of these observations were also true for TSC, except that the DQ and Z were stable within person and behaved more similarly to the standard scores. Further, only DQ and AE had non-homogenous

residual variance, while the norm-referenced scores exhibited excess positive residuals that could threaten model validity.

Discussion

Researchers have a range of options for operationalizing performance on most standardized developmental tests, and the best option must be determined based on the context of use and the intended interpretation of the score. Here, we leveraged data from two GCAND studies to illustrate the analysis of several score types in the longitudinal context, so that we might discuss the quantitative differences as well as the validity case for interpretation of these scores as reflecting individual change. For TSC, we found agreement across norm-referenced scores in that no within-person change was observed, whereas in PTEN, the use of different norm-referenced scores led to different conclusions. However, the floor effects observed for these scores suggest that the results might be biased. In both studies, the absolute scores were consistent with a theoretical developmental curve (faster gains for younger children and slower gains for older children), but both floor and ceiling effects were found for AE. Overall, our results were consistent with our theory-based hypothesis that the person ability score (CSS) would be the most appropriate score type for the longitudinal context.

Model Suitability

We found that the model residuals for all score types were in some way inconsistent with assumptions in the PTEN and/or TSC studies. Future investigators should be aware that analysis of norm-referenced scores is likely to violate assumptions about variance and normality, and that models using RAW and CSS may violate normality assumptions. The violation of normality might in future research be addressable via transformation or the inclusion of additional explanatory variables, but the censoring causing non-homogenous variance in the norm-referenced scores cannot be remediated within the general linear model framework. We selected the multilevel model because it is the standard in the field for modeling an outcome as a function of age, but one might consider a Tobit growth curve (Wang et al., 2008) for data with

high rates of floor effects (such as IQ) or a random effects generalization of quantile regression (Petscher & Logan, 2014) for data where variance increases as a function of scale (such as CSS). These results underscore the importance of reviewing residuals to evaluate the tenability of assumptions in every new model, and suggest that the norm-referenced scores are less well-suited to the standard modeling procedures than the absolute scores.

Validity of Parameter Interpretation

Next, we turn to the validity arguments for interpreting the parameters produced by the statistical models as pertaining to individual change. Some of the current authors (Farmer et al., 2022; Farmer et al., 2023) and others (e.g., Eisengart et al., 2022; Kwok et al., 2022; Shapiro et al., 2024) have argued that for measuring change over time in developmental concepts, the validity case for ability scores like CSS is stronger than that for raw sum scores, AEs, or any type of norm-referenced scores. The most important feature of the ability score with respect to interpretation in longitudinal research is interval-level measurement, because the evaluation of meaningfulness in change rests on the assumption that the meaning is the same regardless of scale location. The AE is an ordinal variable and so it does not meet this standard. As Ostrolenk and Courchesne (2023) pointed out, Wechsler himself lamented that AEs continued to appear in test manuals due to their “firm place in clinical practice … in spite of the fact that these methods violate the philosophy of the Scale” (Wechsler, 1951, p. 381). Dividing an ordinal variable (AE) by an interval-level variable (chronological age) does not yield an interval-level variable, and so this criticism extends to DQ (Ostrolenk & Courchesne, 2023). Further, while the DQ does avoid the normative floor effect, the underlying AE is still subject to the AE floor and ceiling, as observed in the current study. Still, the AE could play an important supporting role in understanding the results of a longitudinal study. The AE can be used to aid clinical interpretation of change in the ability score, which is itself unitless. For example, we understand that the model intercept in this study is the estimated mean score when the average age is 11 years. For TSC, the CSS intercept was 472, which corresponds to an AE of 5 years, 2 months.

For the average 11-year-old TSC participant, the average annualized change of about 4 points per year would translate to an increase to 5 years, 7 months. Note that CSS is the basis of statistical interpretation, and AE is used only as an interpretative support. Ultimately, however, the AE and DQ cannot be reasonably interpreted as an interval level variable and so the valid interpretation of the longitudinal modeling of AE and DQ is simply not possible. AE and DQ should be rejected as endpoints in longitudinal research.

Like AE, the raw score sum is also an ordinal variable, though it has many more levels. However, as observed in the results of this study, the patterns of quantitative results were similar for RAW and CSS. Both exhibited faster change for younger participants that plateaued for older participants, and the model residuals for both were inconsistent with the normality assumption. This was unsurprising, since raw sum scores are an ordered approximation of the ability score (Sijtsma et al., 2024). But when the precise magnitude of change in score is the parameter of interest, the ordinal nature of the raw sum score adversely impacts the validity argument. Importantly, the degree to which this is a problem depends on test construction; where there is more information (items of similar levels of difficulty) on the test, the raw sum score will be closer to interval. Tests developed with considerable resources – like nationally normed and established IQ tests – are likelier to have denser item information than an investigator-created survey instrument. For the SB5, a one-unit change in RAW is equivalent to a one-unit change in CSS for the middle ~50% of the raw sum score range, supporting its interval-level interpretation in that range. As the raw sum score approaches the minimum or maximum extremes, however, this becomes less true (on the SB5, between 2 and 7 RAW points might correspond to a one-unit change in CSS). This could be particularly impactful for studies of individuals with IDD, since samples are likely to have performance in the lower range of scores. Thus, as with any aspect of the validity argument, the extent to which the ordinality of raw sum scores threatens the validity of interpretation must be evaluated for the specific context. Raw sum scores are an absolute measure of ability and the direction of change is

interpretable, but their ordinal nature may adversely impact the valid interpretation of the magnitude of change. Thus, if ability scores are not available for a test, the raw sum score appears to have the next strongest validity argument for use in longitudinal analyses.

Finally, we consider the norm-referenced scores. Unlike RAW, AE, and DQ, normalized standard scores (e.g., IQ, EXIQ) can be considered interval-level to the extent that ability is normally distributed in the population (Eisengart et al., 2022). However, for individuals with IDD, norm-referenced scores like IQ are often significantly limited by floor effects, which the EXIQ is intended to address. As illustrated in both samples in this paper, almost no intermediate values were assigned between the original floor of 40 and the EXIQ floor of 10. While the EXIQ method only moved the standard score floor, the Z-score method did successfully remove it, revealing variability that was censored by the IQ and EXIQ. However, we observed that despite the normative metrics all putatively measuring relative standing of an individual's cognitive ability, the results of statistical analysis did not always lead to the same interpretation. For the PTEN study, the model results for the norm-referenced score types disagreed not only in magnitude but in direction of effect. IQ and EXIQ indicated that on average, skills were gained in a manner consistent with the normative groups (i.e., point estimates for slopes did not differ from zero), but this was not the case for Z, which indicated declines in scores for younger participants. Further, the clinical interpretation of this negative annualized change estimate for Z, like for any norm-referenced score, is indeterminate. Average skill gains may have been slower than necessary to keep up with the normative age tables, there may have been only stability – neither loss nor gain of skills, or perhaps on average, individuals lost skills during study participation. Not only are these three interpretations of Z-score change different from one another, but all of them differ markedly from that of the other norm-referenced scores. Given the psychometric limitations in norm-referenced scores, and the potential for disagreement in conclusions regarding their change, it is difficult to claim that one is more valid than the other.

This is partially driven by the fact that the normative scores measure both change within the person and change in reference groups (each to unknowable degrees), while our desired interpretation corresponds to the direct quantification of individual change as a function of age. This directly threatens the validity of interpreting the results of a longitudinal study as pertaining to change in in the individual. The Projected Retained Ability Score (PRAS) is a new approach to dealing with this limitation (Kronenberger et al., 2021). With PRAS, an individual's performance over time is scaled against the normative group corresponding to that individual's baseline age. We had hoped to include the PRAS in the current study, but the SB5 publisher declined to provide digital versions of the normative scoring tables to facilitate bulk rescoring (ProEd, Personal Communication, 15 November 2023), so we address it only in theory. Indeed, a difference in PRAS does reflect absolute change (like the CSS), but the units are relative (like IQ). In addition to failing to address the significant limitation of standard score floor effects, this could complicate the validity case for interpretation of change scores, depending on the length of follow up of a study (e.g., what is the meaning of comparing change within a now-10-year-old to the baseline distribution of 5-year-olds?). Further theoretical and quantitative evaluation of this method is needed.

Clinical Meaningfulness

Supporting clinical trial readiness is a goal of the Developmental Synaptopathies Consortium, and the results in this paper are intended to aid researchers in the construction of endpoints where a developmental concept is of interest. It is therefore essential to distinguish "clinical meaning," which we have used here to refer to the interpretation of statistical results with respect to the human behavior under study, from clinical *meaningfulness* (Weinfurt, 2019). A statistically detectable difference may not be judged by the patient to be of sufficient magnitude to warrant use of a medication (U.S. Department of Health and Human Services, 2023). Thus, regulatory agencies require both statistical evidence of efficacy and qualitative evidence supporting the clinical meaningfulness of that statistical effect to stakeholders (U.S.

Department of Health and Human Services, 2023). In this study, we have focused only on the former. The meaningfulness of an effect depends on the context, and so one must determine via qualitative methods what a clinically meaningful effect is for an individual study, regardless of the selected scoring method. We have sometimes encountered the argument that the clinical meaningfulness of the score types mentioned here can be established by comparing to their SD or SEM. This is especially true of norm-referenced scores, for which meaningful change is often colloquially defined based on the population standard deviation (e.g., one-half an SD). This is a distribution-based method of establishing meaningful differences, which is not considered adequate by regulatory agencies (U.S. Department of Health and Human Services, 2022). The anchor-based method, though not without limitations (Wyrwich & Norman, 2023), is one alternative approach. This method compares quantitative change on the outcome measure against qualitative ratings of improvement (usually from the patient or caregiver's perspective) (e.g., Chatham et al., 2018). And so, while we cannot speak to the clinical meaningfulness of any of the statistical results described in this study because we have not performed the necessary qualitative work, we refer interested readers to work exploring methods for establishing clinically meaningful change for individuals with GCAND and related conditions (Duong et al., 2021).

Conclusion

Whether an endpoint is fit-for-purpose depends on the context of use, and in the case of longitudinal research, the goal is to derive information about individual change in ability as a function of time. Here, we described and illustrated theoretical and quantitative threats to validity for several types of scores from a standardized developmental test. Some limitations of the norm-referenced scores, such as floor effects, were observable in the data. However, other limitations, such as the indeterminacy of change in norm-referenced scores and the ordinal nature of AEs and raw sum scores, are not observable in model results and must be considered theoretically. Researchers must consider the validity of each score type for their particular

context. Based on theory and the statistical results of this study, we argue that for longitudinal studies of people with IDD, the person ability score is most appropriate.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bishop, S. L., Farmer, C., & Thurm, A. (2015). Measurement of nonverbal IQ in autism spectrum disorder: scores in young adulthood compared to early childhood. *J Autism Dev Disord*, *45*(4), 966-974. https://doi.org/10.1007/s10803-014-2250-3

Busch, R. M., Frazier Ii, T. W., Sonneborn, C., Hogue, O., Klaas, P., Srivastava, S., Hardan, A. Y., Martinez-Agosto, J. A., Sahin, M., & Eng, C. (2023). Longitudinal neurobehavioral profiles in children and young adults with PTEN hamartoma tumor syndrome and reliable methods for assessing neurobehavioral change. *J Neurodev Disord*, *15*(1), 3. https://doi.org/10.1186/s11689-022-09468-4

Busch, R. M., Srivastava, S., Hogue, O., Frazier, T. W., Klaas, P., Hardan, A., Martinez-Agosto, J. A., Sahin, M., & Eng, C. (2019). Neurobehavioral phenotype of autism spectrum disorder associated with germline heterozygous mutations in PTEN. *Transl Psychiatry*, *9*(1), 253. https://doi.org/10.1038/s41398-019-0588-1

Chatham, C. H., Taylor, K. I., Charman, T., Liogier D'ardhuy, X., Eule, E., Fedele, A., Hardan, A. Y., Loth, E., Murtagh, L., Del Valle Rubido, M., San Jose Caceres, A., Sevigny, J., Sikich, L., Snyder, L., Tillmann, J. E., Ventola, P. E., Walton-Bowen, K. L., Wang, P. P., Willgoss, T., & Bolognani, F. (2018). Adaptive behavior in autism: Minimal clinically important differences on the Vineland-II. *Autism Res*, *11*(2), 270-283. https://doi.org/10.1002/aur.1874

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual review of psychology*, *62*, 583-619.

Duong, T., Staunton, H., Braid, J., Barriere, A., Trzaskoma, B., Gao, L., Willgoss, T., Cruz, R., Gusset, N., Gorni, K., Randhawa, S., Yang, L., & Vuillerot, C. (2021). A Patient-Centered Evaluation of Meaningful Change on the 32-Item Motor Function Measure in Spinal Muscular Atrophy Using Qualitative and Quantitative Data. *Front Neurol*, *12*, 770423. https://doi.org/10.3389/fneur.2021.770423

Eisengart, J. B., Daniel, M. H., Adams, H. R., Williams, P., Kuca, B., & Shapiro, E. (2022). Increasing precision in the measurement of change in pediatric neurodegenerative disease. *Molecular Genetics and Metabolism*, *137*(1), 201-209. https://doi.org/https://doi.org/10.1016/j.ymgme.2022.09.001

Farmer, C., Kaat, A. J., Berry-Kravis, E., & Thurm, A. (2022). Chapter One - Psychometric perspectives on developmental outcome and endpoint selection in treatment trials for genetic conditions associated with neurodevelopmental disorder. In A. J. Esbensen & E. K. Schworer (Eds.), *International Review of Research in Developmental Disabilities* (Vol. 62, pp. 1-39). Academic Press. https://doi.org/https://doi.org/10.1016/bs.irrdd.2022.05.001

Farmer, C., Thurm, A., Troy, J. D., & Kaat, A. J. (2023). Comparing ability and norm-referenced scores as clinical trial outcomes for neurodevelopmental disabilities: a simulation study. *J Neurodev Disord*, *15*(1), 4. https://doi.org/10.1186/s11689-022-09474-6

Hessl, D., Nguyen, D. V., Green, C., Chavez, A., Tassone, F., Hagerman, R. J., Senturk, D., Schneider, A., Lightbody, A., Reiss, A. L., & Hall, S. (2009). A solution to limitations of cognitive testing in children with intellectual disabilities: the case of fragile X syndrome. *J Neurodev Disord*, *1*(1), 33-45. https://doi.org/10.1007/s11689-008-9001-8

Kronenberger, W. G., Harrington, M., & Yee, K. S. (2021). Projected Retained Ability Score (PRAS): A New Methodology for Quantifying Absolute Change in Norm-Based Psychological Test Scores Over Time. *Assessment*, *28*(2), 367-379. https://doi.org/10.1177/1073191119872250

Kwok, E., Feiner, H., Grauzer, J., Kaat, A., & Roberts, M. Y. (2022). Measuring Change During Intervention Using Norm-Referenced, Standardized Measures: A Comparison of Raw Scores, Standard Scores, Age Equivalents, and Growth Scale Values From the Preschool Language Scales-Fifth Edition. *J Speech Lang Hear Res*, *65*(11), 4268-4279. https://doi.org/10.1044/2022_jslhr-22-00122

Levy, T., Foss-Feig, J. H., Betancur, C., Siper, P. M., Trelles-Thorne, M. D. P., Halpern, D., Frank, Y., Lozano, R., Layton, C., Britvan, B., Bernstein, J. A., Buxbaum, J. D., Berry-Kravis, E., Powell, C. M., Srivastava, S., Sahin, M., Soorya, L., Thurm, A., & Kolevzon, A. (2022). Strong evidence for genotype-phenotype correlations in Phelan-McDermid syndrome: results from the developmental synaptopathies consortium. *Hum Mol Genet*, *31*(4), 625-637. https://doi.org/10.1093/hmg/ddab280

Ostrolenk, A., & Courchesne, V. (2023). Examining the validity of the use of ratio IQs in psychological assessments. *Acta Psychol (Amst)*, *240*, 104054. https://doi.org/10.1016/j.actpsy.2023.104054

Petscher, Y., & Logan, J. A. R. (2014). Quantile regression in the study of developmental sciences. *Child Dev*, *85*(3), 861-881. https://doi.org/10.1111/cdev.12190

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition, Technical Manual*. Riverside Publishing.

Sanders, S. J., Sahin, M., Hostyk, J., Thurm, A., Jacquemont, S., Avillach, P., Douard, E., Martin, C. L., Modi, M. E., Moreno-De-Luca, A., Raznahan, A., Anticevic, A., Dolmetsch, R., Feng, G., Geschwind, D. H., Glahn, D. C., Goldstein, D. B., Ledbetter, D. H., Mulle, J. G., Pasca, S. P., Samaco, R., Sebat, J., Pariser, A., Lehner, T., Gur, R. E., & Bearden, C. E. (2019). A framework for the investigation of rare genetic disorders in neuropsychiatry. *Nat Med*, *25*(10), 1477-1487. https://doi.org/10.1038/s41591-019-0581-5

Sansone, S. M., Schneider, A., Bickel, E., Berry-Kravis, E., Prescott, C., & Hessl, D. (2014). Improving IQ measurement in intellectual disabilities using true deviation from population norms. *Journal of Neurodevelopmental Disorders*, *6*(1), 16. https://doi.org/10.1186/1866-1955-6-16

Shapiro, E. G., Eisengart, J. B., Whiteman, D., & Whitley, C. B. (2024). Ability change across multiple domains in mucopolysaccharidosis (Sanfilippo syndrome) type IIIA. *Mol Genet Metab*, *141*(2), 108110. https://doi.org/10.1016/j.ymgme.2023.108110

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024). Recognize the Value of the Sum Score, Psychometrics' Greatest Accomplishment. *Psychometrika*, 1-34.

Soorya, L., Leon, J., Trelles, M. P., & Thurm, A. (2018). Framework for assessing individuals with rare genetic disorders associated with profound intellectual and multiple disabilities (PIMD): the example of Phelan McDermid Syndrome. *Clin Neuropsychol*, *32*(7), 1226-1255. https://doi.org/10.1080/13854046.2017.1413211

Timmerman, M. E., Voncken, L., & Albers, C. J. (2021). A tutorial on regression-based norming of psychological tests with GAMLSS. *Psychological Methods*, *26*(3), 357-373. https://doi.org/10.1037/met0000348

U.S. Department of Health and Human Services. (2022). Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments [Guidance Document]. 40.

U.S. Department of Health and Human Services. (2023). Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate behavioral research*, *43*(3), 476-496.

Wechsler, D. (1951). Equivalent test and mental ages for the WISC. *Journal of consulting psychology*, *15*(5), 381.

Weinfurt, K. P. (2019). Clarifying the Meaning of Clinically Meaningful Benefit in Clinical Research: Noticeable Change vs Valuable Change. *JAMA*, *322*(24), 2381-2382. https://doi.org/10.1001/jama.2019.18496

Wright, C. F., FitzPatrick, D. R., & Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet*, *19*(5), 253-268. https://doi.org/10.1038/nrg.2017.116

Wyrwich, K. W., & Norman, G. R. (2023). The challenges inherent with anchor-based approaches to the interpretation of important change in clinical outcome assessments. *Qual Life Res*, *32*(5), 1239-1246. https://doi.org/10.1007/s11136-022-03297-7

Figure Legends

Figure 1. Observed data by age for each study. IQ = intelligence quotient; EXIQ = extended IQ; Z = Z-normalized score; DQ = developmental quotient; AE = age equivalent; CSS = change sensitive score. Each observation is marked with a filled circle and observations from the same individual are connected by a solid line. For IQ, EXIQ, Z, and DQ, the dotted line reflects the expected population average. For RAW, the dotted line indicates the raw score corresponding to AEs plotted across the X axis. For AE, the dotted line indicates the AE corresponding to each chronological age. There are no normative values for CSS, but the CSS corresponding to each AE is plotted (AE is on the chronological age axis).

Figure 2. Distributions and interrelationships of each score type at baseline. IQ = intelligence quotient; EXIQ = extended IQ; Z = Z-normalized score; DQ = developmental quotient; AE = age equivalent; CSS = change sensitive score. Panel A: PTEN cohort data. Panel B: TSC cohort data. Both panels: Diagonal is the density plot per score. Below the diagonal is scatter plot of scores on X and Y axis. Above the diagonal is Spearman rank-order correlation for scores on X and Y axis.

Figure 3. Annualized change estimates. IQ = intelligence quotient; EXIQ = extended IQ; Z = Z-normalized score; DQ = developmental quotient; AE = age equivalent; CSS = change sensitive score. Both panels: Contrasts were used to generate the predicted fixed estimate for DURATION (annualized change) at several hypothetical ages. This can be expressed as a function of $\overline{AGE}$ (see Table 4 for DURATION* $\overline{AGE}$ term), though if the interaction was not different from zero the DURATION estimate will be similar across values of $\overline{AGE}$. Panel A: Results for the PTEN sample. Panel B: Results for the TSC sample.

Tables

Table 1. Summary of available methods for operationalizing performance on the SB5 Full Scale Composite.

Feature	Relative				Absolute			Hybrid
	Intelligence Quotient (IQ)	***Extended IQ (EXIQ)***	***Deviation Z Score (Z)***	***Developmental Quotient (DQ)***	***Raw Sum Score***	***Age Equivalent (AE)***	***Change Sensitive Score (CSS)***	***Projected Retained Ability Score (PRAS)***
Possible Range on SB5	40 – 160	10 – 225	-17.5 – 231.7	0 – 1050	0 – 358	24 – 252	376 – 592	40 – 160
Floor types	Test, normative	Test, normative	Test	Test, age equivalent	Test	Test, age equivalent	Test	Test, normative
Derivation	Sums of normalized scaled scores are tabulated and smoothed within age groups.	Distance of the CSS from the normative age group mean in standard deviation units; used only when IQ is 40 or 160.	Distance of the summed subtest raw scores from the normative age group mean in standard deviation units.	AE is divided by chronological age.	Sum of item-level scores.	Small-range age groupings formed and mean CSS scores plotted; age grouping corresponding to CSS obtained from best-fitting regression lines.	Ability estimates from the Rasch model converted to CSS metric (centered at 500 for AE of 10:00).	Sums of normalized scaled scores are tabulated and smoothed within age groups.
Measurement level	Interval (if underlying ability is normally distributed)	Interval (if underlying ability is normally distributed)	Interval if raw score is normally distributed within the normative group (typically is not); in practice treated as interval.	Undefined but in practice treated as interval.	(poly) Ordinal	Ordinal	Interval	Interval (if underlying ability is normally distributed)
Units	Standard deviations	Standard deviations	If raw score is not normally distributed in normative group, does not have quantitative units but in practice, treated as	In practice but without empirical or theoretical basis, treated as standard deviation units.	No quantitative units (ordinal)	No quantitative units (ordinal). AEs are labeled as months of age but these are category labels, not continuous units	Quantitative but units have no meaningful scale (latent score)	Standard deviations

			standard deviation units. If the underlying distribution is not normal, Z-scores do not correspond to percentiles.			of measurement.		
Intended context of use	Diagnostic	Diagnostic	Diagnostic	Diagnostic	Descriptive	Descriptive	Change monitoring	Change monitoring
Interpretation of a single score with respect to measured construct	The relative standing in a normal distribution of their peers' performance.	The relative standing in a normal distribution of their peers' performance.	Distance from the norm-group mean, but if underlying distribution is normal then relative standing in a normal distribution of their peers' performance.	In practice but without empirical or theoretical basis, it is interpreted in the same way as IQ.	Number of items passed	The age group whose average score is equal to the observed performance.	Amount of the measured construct demonstrated by the individual.	The relative standing in a normal distribution of their baseline age-peers' performance.
Interpretation of decrease in score from T1 to T2 with respect to measured construct	Indeterminate	Indeterminate	Indeterminate	Indeterminate	Ability declined	Ability declined	Ability declined	Ability declined

Note: Inspired by Eisengart et al. (2022), we prepared a summary of relevant characteristics for each score type on the SB5. The ranges provided pertain only to the Full Scale composite of the SB5, but the remaining columns should be generally applicable to the same scores from other composites or other tests. The PRAS was not evaluated in this study because it was not available.

Table 2. Baseline characteristics of cohorts used in analysis

		PTEN	TSC
N		91	81
Gender	Female, n (%)	24 (26%)	32 (40%)
	Male, n (%)	67 (74%)	49 (60%)
Race	American Indian, n (%)	0 (0%)	0 (0%)
	Asian, n (%)	7 (8%)	1 (1%)
	Black, n (%)	2 (2%)	3 (4%)
	Multiple, n (%)	13 (14%)	5 (6%)
	Pacific Islander, n (%)	1 (1%)	0 (0%)
	White, n (%)	64 (70%)	70 (86%)
	Unknown or not reported, n (%)	4 (4%)	2 (2%)
Ethnicity	Hispanic, Latino, or Spanish origin, n (%)	9 (10%)	15 (19%)
	Not Hispanic, Latino, or Spanish origin, n (%)	80 (88%)	64 (79%)
	Unknown or not reported, n (%)	2 (2%)	2 (2%)
Initial age (years)	Median [IQR]	9.37 [6.42, 13.39]	8.93 [5.25, 12.54]
	Range	3.56 – 21.99	3.00 – 20.30
Initial SB5 Full Scale Score	IQ, mean (SD)	76.88 (26.67)	61.47 (17.98)
	EXIQ, mean (SD)	73.59 (32.32)	57.40 (24.64)
	Z, mean (SD)	78.34 (29.12)	65.01 (22.82)
	DQ, mean (SD)	73.21 (38.83)	53.84 (21.26)
	RAW, mean (SD)	137.93 (66.07)	101.72 (55.29)
	AE, mean (SD)	83.20 (54.07)	56.58 (33.45)
	CSS, mean (SD)	474.60 (26.15)	460.83 (23.29)

Note: SB5 = Stanford Binet, 5th Edition; SD = standard deviation; IQ = intelligence quotient; EXIQ = extended IQ; Z = Z-normalized score; DQ = developmental quotient; RAW = raw sum score; AE = age equivalent; CSS = change sensitive score. The initial visit is the first visit with an SB5, which was not necessarily the first study visit. An insufficient number of individuals in the Phelan-McDermid study were able to take the SB5 (i.e., the test floor was too high) and the cohort was therefore not included in analysis.

Table 3. Fixed effects from hierarchical linear models.

Study	Score	Parameter	Intercept	$\overline{AGE}$	DURATION	$\overline{AGE}^2$	DURATION* $\overline{AGE}$
PTEN	IQ	Est [95% CI]	77.45 [69.73, 85.16]	-0.92 [-2.29, 0.44]	-0.06 [-1.04, 0.92]	-0.01 [-0.23, 0.21]	0.16 [-0.02, 0.34]
		t (p-value)	t(88.4)=19.68 (<.001)	t(88.6)=-1.33 (0.187)	t(59.6)=-0.12 (0.906)	t(87.8)=-0.1 (0.92)	t(59.2)=1.75 (0.085)
	EXIQ	Est [95% CI]	73.56 [64.25, 82.88]	-1.25 [-2.9, 0.4]	-0.87 [-2.24, 0.51]	0 [-0.26, 0.26]	0.22 [-0.04, 0.47]
		t (p-value)	t(90.2)=15.48 (<.001)	t(90.2)=-1.49 (0.14)	t(65.4)=-1.24 (0.221)	t(86.9)=0.01 (0.994)	t(62.5)=1.65 (0.105)
	Z	Est [95% CI]	76.5 [68.18, 84.83]	-1.35 [-2.82, 0.12]	-1.24 [-2.41, -0.06]	0.06 [-0.18, 0.29]	0.27 [0.05, 0.49]
		t (p-value)	t(89.6)=18.01 (<.001)	t(89.6)=-1.79 (0.076)	t(66.5)=-2.07 (0.043)	t(86.9)=0.48 (0.634)	t(63.5)=2.43 (0.018)
	DQ	Est [95% CI]	75.37 [64.88, 85.87]	-2.12 [-3.98, -0.27]	-1.63 [-3.18, -0.08]	-0.11 [-0.41, 0.19]	0.06 [-0.23, 0.35]
		t (p-value)	t(87.8)=14.08 (<.001)	t(88.2)=-2.25 (0.027)	t(76.7)=-2.06 (0.043)	t(87.7)=-0.69 (0.49)	t(76.1)=0.41 (0.68)
	RAW	Est [95% CI]	167.66 [150.3, 185.01]	7.51 [4.44, 10.58]	9.55 [7.53, 11.57]	-0.94 [-1.43, -0.45]	-0.98 [-1.36, -0.6]
		t (p-value)	t(89.6)=18.93 (<.001)	t(89.7)=4.8 (<.001)	t(66.5)=9.28 (<.001)	t(87.4)=-3.76 (<.001)	t(63.9)=-5.1 (<.001)
	AE	Est [95% CI]	101.74 [87.11, 116.36]	6.02 [3.43, 8.61]	7.23 [5.14, 9.32]	-0.54 [-0.95, -0.13]	-0.58 [-0.97, -0.19]
		t (p-value)	t(91.2)=13.63 (<.001)	t(91.2)=4.55 (<.001)	t(67.4)=6.77 (<.001)	t(89.4)=-2.56 (0.012)	t(67.2)=-2.92 (0.005)
	CSS	Est [95% CI]	485.9 [479.03, 492.76]	2.74 [1.53, 3.95]	3.6 [2.71, 4.48]	-0.36 [-0.55, -0.16]	-0.38 [-0.54, -0.21]
		t (p-value)	t(88)=138.73 (<.001)	t(88.2)=4.43 (<.001)	t(68.6)=7.96 (<.001)	t(87.9)=-3.59 (0.001)	t(65.5)=-4.49 (<.001)
TSC	IQ	Est [95% CI]	59.59 [54.16, 65.02]	-1.12 [-1.97, -0.28]	0.58 [-0.48, 1.63]	0.09 [-0.09, 0.26]	0.1 [-0.14, 0.34]
		t (p-value)	t(77.4)=21.52 (<.001)	t(77.5)=-2.61 (0.011)	t(47.5)=1.07 (0.29)	t(76.9)=0.96 (0.34)	t(53.3)=0.78 (0.438)
	EXIQ	Est [95% CI]	54.61 [47.51, 61.71]	-1.52 [-2.62, -0.41]	1.1 [-0.47, 2.68]	0.13 [-0.1, 0.35]	-0.16 [-0.52, 0.21]
		t (p-value)	t(76.7)=15.08 (<.001)	t(77.7)=-2.69 (0.009)	t(108.5)=1.37 (0.172)	t(77.2)=1.09 (0.278)	t(109.2)=-0.85 (0.396)
	Z	Est [95% CI]	58.07 [51.5, 64.65]	-1.29 [-2.31, -0.26]	-0.29 [-1.42, 0.84]	0.29 [0.08, 0.5]	0.19 [-0.07, 0.45]
		t (p-value)	t(78.1)=17.31 (<.001)	t(77.8)=-2.46 (0.016)	t(53.4)=-0.5 (0.62)	t(77.3)=2.69 (0.009)	t(58.5)=1.41 (0.165)
	DQ	Est [95% CI]	50.67 [44.62, 56.72]	-1.96 [-2.9, -1.02]	-0.9 [-2.09, 0.28]	0.07 [-0.13, 0.26]	-0.13 [-0.4, 0.14]
		t (p-value)	t(76.2)=16.41 (<.001)	t(76.9)=-4.09 (<.001)	t(56.3)=-1.49 (0.141)	t(76.7)=0.69 (0.494)	t(60.8)=-0.93 (0.357)
	RAW	Est [95% CI]	127.92 [114.52, 141.33]	6.63 [4.51, 8.75]	9.06 [6.71, 11.42]	-0.58 [-1, -0.16]	-1.42 [-1.96, -0.88]
		t (p-value)	t(79.9)=18.7 (<.001)	t(77.1)=6.13 (<.001)	t(58.2)=7.55 (<.001)	t(76)=-2.7 (0.008)	t(63.3)=-5.19 (<.001)
	AE	Est [95% CI]	68.21 [59.89, 76.54]	3.89 [2.52, 5.26]	4.74 [3.41, 6.06]	-0.2 [-0.45, 0.05]	-0.4 [-0.7, -0.1]
		t (p-value)	t(82.4)=16.06 (<.001)	t(76.5)=5.57 (<.001)	t(49.4)=7.02 (<.001)	t(60.2)=-1.58 (0.119)	t(55.2)=-2.59 (0.012)
	CSS	Est [95% CI]	471.62 [465.98, 477.26]	2.72 [1.84, 3.6]	3.72 [2.75, 4.68]	-0.24 [-0.42, -0.06]	-0.69 [-0.91, -0.47]

	t (p-value)	t(78.3)=163.93 (<.001)	t(77.1)=6.05 (<.001)	t(55.6)=7.53 (<.001)	t(76.5)=-2.61 (0.011)	t(60.4)=-6.12 (<.001)

Note: IQ = intelligence quotient; EXIQ = extended IQ; Z = Z-normalized score; DQ = developmental quotient; RAW = raw sum score; AE = age equivalent; CSS = change sensitive score; Est = parameter estimate; CI = confidence interval. $\overline{AGE}$ was centered at 11 years. The statistical interpretation of each parameter is described in the methods section. Model results are summarized in Table 4. Random effects are available at [SEE SUPPLEMENTARY MATERIALS PROVIDED FOR REVIEW].

Table 4. Summary of hierarchical linear model results.

Score	Floor effects	Residual Diagnostics		Annualized (Within-person) Change	Point-in-time Estimated Score (Between-person differences)	Cohort Effect
		Variance	*Normality*	*DURATION and DURATION* $\overline{AGE}$*	$\overline{AGE}$ *and* $\overline{AGE}^2$	
PTEN						
IQ	11% obs, 14% sample	Conical	OK	Stable across ages	No differences across ages	No
EXIQ	10% obs, 12% sample	Conical	OK	Stable across ages	No differences across ages	No
Z	No	OK	OK	Declines for younger, stability for older	No differences across ages	Yes
DQ	See AE	Conical	OK	Stable across ages	Older < younger	No
RAW	No	OK	Excess positive	Gains for younger, stability for older	Older > younger	No
AE	6% obs, 9% sample; Ceiling: 7% obs, 7% sample	Conical	OK	Gains for younger, stability for older	Older > younger	No
CSS	No	OK	Excess positive	Gains for younger, stability for older	Older > younger	No
TSC						
IQ	12% obs, 17% sample	OK	Excess positive	Stable across ages	Older < younger	Yes
EXIQ	12% obs, 17% sample	OK	Excess positive	Stable across ages	Older < younger	Yes
Z	No	OK	Excess positive	Stable across ages	Older < younger	No
RAW	No	OK	Excess positive	Gains for younger, stability for older	Older > younger	No
DQ	See AE	Conical	OK	Stable across ages	Older < younger	No
AE	7% obs, 12% sample; Ceiling: <1% obs, 1% sample	Conical	OK	Gains for younger, stability for older	Older > younger	No
CSS	No	OK	OK	Gains for younger, stability for older	Older > younger	No

Note: IQ = intelligence quotient; EXIQ = extended IQ; Z = Z-normalized score; DQ = developmental quotient; RAW = raw sum score; AE = age equivalent; CSS = change sensitive score; obs = total number of observations. Floor effects are described as a proportion of the total observations (% obs) and the proportion of individuals with at least one censored value (% sample). These floor effects refer only to standard score and AE floor, not to the test floor (i.e., participants who could not take the test). Cohort effect is marked "Yes" when the confidence interval for the within- and between-subject effects excluded one another, "No" when they did not. While no raw scores at the floor were observed, we note that if a participant received a score of zero on several subtests, the testing would

have been halted. Annualized change is summarized as “stable” if point estimates generally did not differ from zero, but trends toward negative or positive slopes can be observed in the figure.