JOURNAL OF CHILD AND ADOLESCENT PSYCHOPHARMACOLOGY
Volume 00, Number 00, 2025

Pp. 000–000
DOI: 10.1089/cap.2024.0107

Open camera or QR reader and scan code to access this article and other resources online.

Vineland-3 Growth Scale Values: Psychometric Properties for Clinical Trial Readiness in SCN2A

Aaron J. Kaat, PhD,[1] Lindsey Evans, PhD,[2] Amanda N. Nili, PhD,[1] Katherine Paltell, PhD,[3] Arielle Kaiser, PhD,[3] Erica Anderson, PhD,[4] Leah Schust Myers,[5] and Anne T. Berg, PhD[1,5]

Abstract

Purpose: The Vineland Adaptive Behavior Scales—3rd Edition (Vineland-3) is one of the most used measures of adaptive behavior among those with sodium channel protein type 2 subunit alpha related disorders (SCN2A-RDs). Several disease-modifying treatments are in early trials for SCN2A-RDs, and as such, clinical outcome assessments (COAs) are necessary. The Vineland-3 introduced growth scale values (GSVs), which are useful for measuring within-person change and thus may be useful in future clinical trials. The purpose of this study was to evaluate the psychometric properties of the Vineland-3 GSVs in SCN2A-RDs in preparation for future clinical trials.

Methods: A sample of 65 individuals with SCN2A-RDs (mean = 108, SD = 76.0 months) was recruited for a clinical trial readiness study. The Vineland-3 Comprehensive Interview was administered by trained raters at regular intervals. Multiple psychometric properties were evaluated, including floor and ceiling effects, split-half internal consistency, test-retest reliability, and inter-rater reliability (on approximately 20% of all completions).

Results: Floor effects were relatively infrequent on the GSV metric but occurred on all subdomains using the norm-referenced v-scale metric. Split-half and test-retest reliability were excellent for all subdomains (r_{xx} >0.95 and inter-class correlation coefficient [ICC] >0.90, respectively), except for coping, which still maintained adequate reliability (r_{xx} = 0.87, ICC = 0.65). Inter-rater reliability was also very strong, though it was more variable (α_{kripp} range 0.78–1.00).

Conclusion: The Vineland-3 holds great potential as a COA in SCN2A-RDs; it exhibited very strong psychometric properties in this sample. This is a prerequisite level of evidence needed to demonstrate that a measure is fit-for-purpose for future clinical trials. While some reliability was high, some domains (e.g., domestic) still exhibited problems related to floor effects, which may suggest that they are less relevant to this population. Future studies should expand on this with mixed-methods research for prioritizing concepts of interest on the Vineland-3.

Keywords: SCN2A, Vineland-3, psychometrics, adaptive behavior

Introduction

Sodium channel protein type 2 subunit alpha related disorders (SCN2A-RDs) are caused by pathogenic variants in the *SCN2A* gene, which codes for the voltage-gated sodium channel subunit alpha $Na_v1.2$. SCN2A-RDs are rare, with population estimates of approximately 1 in 100,000 live births (Symonds et al., 2019; Wolff et al., 2017). The original reports associating SCN2A with epilepsy were for a familial syndrome, self-limited familial neonatal and infantile epilepsy previously known as benign familial neonatal or infantile epilepsy (Heron et al., 2002). Since then, variants in SCN2A have been associated with a much more severe spectrum of neurodevelopmental disorders characterized by intellectual disability, severe developmental and epileptic encephalopathy, and autism spectrum disorder (Wolff et al., 2017). Individuals with SCN2A-RD have a range of other morbidities, including movement disorders, dysautonomias, cortical visual impairment, and other neurological and nonneurological conditions (Sanders et al., 2018; Berg et al., 2021).

Monogenic conditions, such as SCN2A-RDs, are prime targets for disease-targeted (or "precision") therapies. Several such pharmacological interventions are currently in early-stage trials. Although seizures are an important clinical feature for most individuals with

[1]Northwestern University Feinberg School of Medicine, Chicago, IL, USA.
[2]Illinois Institute of Technology, Chicago, IL, USA.
[3]University of Illinois at Chicago, Chicago, IL, USA.
[4]Independent Practice, Evanston, IL, USA.
[5]FamilieSCN2A Foundation, Gettysburg, PA, USA.

SCN2A-RDs, they are not always the most important clinical endpoint, and, in some individuals, they may not even be present. This heterogeneity in presentation creates challenges as nonseizure outcomes must be identified, defined, and measured to identify trial endpoints. The importance of delineating such endpoints lies in the need to assess and intervene on the comprehensive phenotype. In this way, clinical trials can ensure that the therapeutic benefit is not solely affecting one clinical feature by instead reflecting the concepts of interest prioritized by patients and families and are inclusive of the heterogeneity within SCN2A-RDs. Individuals with SCN2A-RDs have severe to profound impairments that render typical clinical outcome assessments (COA) inappropriate for use. A recent analysis of the Simons Foundation Autism Research Initiative demonstrated in a cohort of 64 children and young adults with SCN2A-RDs that standardized scores on the Vineland Adaptive Behavior Scales-II declined with age, and severe floor effects were evident, even for the raw scores, in many of the domains and subdomains of the instrument (Berg et al., 2021). This decline with age and the floor effects have been seen in other diseases where affected individuals have severe to profound impairments (Yang et al., 2016; Semmel et al., 2019).

Alternatives to using an instrument as initially intended with norm-referenced scoring include use of alternative scoring. The Vineland Adaptive Behavioral Scales (both Versions II and 3) is an observer-reported measure usually completed by parents, either as a caregiver-report measure or as part of a clinical interview, that assesses adaptive behavior in a set of fundamental domains, motor, communication, sociability, and daily living skills. It has been well-accepted and often used in research, including clinical trials (Devinsky et al., 2017; Devinsky et al., 2018; O'Callaghan et al., 2011).

Increasingly, randomized trials are targeting specific rare diseases and syndromes. Apart from seizures, which are relatively easy to ascertain, selection of appropriate outcome domains, measures, and endpoints has been a challenge that these trials must meet in order to provide an effective evaluation of treatment impact.

One domain that is especially relevant for clinical trials within SCN2A is adaptive behavior. Indeed, adaptive behavior is often used as a proxy of overall development and cognitive functioning among individuals with severe or profound intellectual disability. The American Association on Intellectual and Developmental Disabilities (AAIDD) defines adaptive behavior as "the collection of conceptual, social, and practical skills that have been learned and are performed by people in their everyday lives'' (Schalock et al., 2010; Tassé et al., 2012). Adaptive behavior, then, is determined partially based upon age-appropriate expectations, and thus they change across the age span. This definition has largely been adopted by other groups, including (with slight modifications) within the Diagnostic and Statistical Manual 5th Edition (APA, 2013). One of the most commonly used adaptive behavior assessments is the Vineland Adaptive Behavior Scales—3rd Edition (Vineland) (Sparrow et al., 2016). The Vineland uses a slightly different definition of adaptive behavior—specifically that it is "the performance of daily activities required for personal and social sufficiency" (Sparrow et al., 2016). It conceptualizes adaptive behavior as having domains for communication, socialization, daily living skills, and motor, and 11 subdomains within these higher-order domains. This definition is not contradictory to the AAIDD definition (Schalock et al., 2010) but rather complements it. Self-sufficiency is seen as key to social competence and appropriate functioning.

The Vineland is commonly included within clinical trials. It serves multiple purposes depending on the trial, including as an eligibility criterion, as a method to characterize the sample at baseline, or as an outcome at the end of the trial. However, the Food and Drug Administration (FDA) has standards for trial outcomes. Following the passage of the 21st Century Cures Act (Hudson and Collins, 2017), there has been a growing emphasis within the FDA on patient-focused drug development (Basch et al., 2015; Perfetto et al., 2015). Various guidance documents specifically address rare diseases (ADMINISTRATION, U.S. FOOD AND DRUG [FDA], 2018, 2019), gathering input on patients' priorities (FDA, 2022a), selecting clinical outcome assessments (COAs; FDA, 2022b), and most recently, how to incorporate COAs into clinical endpoints (FDA, 2023). For the Vineland-3 (or really any outcome measure) to be considered fit-for-purpose and acceptable for use as a COA in a clinical trial, several psychometric criteria must be met. These include adequate internal consistency, good inter- and intra-rater reliability, absence of substantial floor or ceiling effects, ability to discriminate between different health states, and sensitivity to meaningful change over time. As part of the *SCN2A* Clinical Trials Readiness Study (CTRS), we took the perspective of patient-focused COA assessment to evaluate the psychometric adequacy of the Vineland-3. It is a meaningful potential outcome measure for precision therapy trials for SCN2A-RDs, and therefore establishing evidence aligned with the FDA guidance documents is of critical importance.

Methods

Participants

The SCN2A-CTRS was community-based participatory research, where the study was funded by the FamilieSCN2A Foundation, which helped design the study protocol and support participant accrual. A total of 65 individuals from 7 countries participated in this study. Parents had to have adequate English abilities to respond to survey measures and complete semi-structured interviews in English, and their children (the "participants") had to be at least 1 year old at the time of study entry. Participant demographic and clinical information is provided in Table 1. Parents were respondents for their children, including adults with intellectual disabilities. The SCN2A-CTRS was performed primarily as a web-based survey designed and administered in CLIRINX©. The survey included questions that allowed the determination of basic functional abilities including mobility, communication, hand use (any purposeful grasp), and eating ability (exclusively G-tube fed or not) (Paulson and Vargus-Adams, 2017; Hidecker et al., 2011; Towns et al., 2018). The sample exhibited significant functional impairments on all measures, as shown in Table 1, and was consistent with previously published descriptions of SCN2A-RDs (e.g., Wolff et al., 2017; Sanders et al., 2018).

Potential COAs

The primary COA we evaluated was the Vineland-3 comprehensive interview (Sparrow et al., 2016), which is a semi-structured parent/caregiver interview designed to assess adaptive behavior across the lifespan. The Vineland provides an overall Adaptive Behavior Composite (ABC) score, which is derived from the communication, socialization, and daily living skills domains; the motor domain was also administered to all participants regardless of age. Each domain is further comprised of two or three subdomains for a

Table 1. Participant Demographic and Clinical Information

Characteristic	*Mean*	*SD*
Chronological age in months	108	76.0
Vineland-3 Adaptive Behavior Composite	34.3	12.7
Number of antiseizure medications ($n = 56$)	Median = 2	IQR = 1 to 3.25

Category	*Characteristic*	N	%
Sex assigned at birth	Male	37	56.9%
	Female	28	43.1%
Co-occurring clinical phenotypes	Epilepsy	56	86.2%
	Autism Spectrum Disorder	34	52.3%
	Cortical/Cerebral Visual Impairment	29	44.6%
	Lack of symbolic communication	30	46.1%
Recent seizure activity	Within the last week	25	38.5%
	More than 1 week but less than a month	4	6.2%
	More than a month but less than 3 months ago	4	6.2%
	More than 3 months ago	23	35.4%
	No epilepsy reported	9	13.8%

total of 11 subdomains. Most subdomains begin at birth, but the coping subdomain begins at 2 years, and the written, domestic, and community subdomains begin at 3 years. The subdomains within motor do not provide norm-referenced scores after 10 years old.

The comprehensive interview form starts based on age or expected developmental level. The raw score is the sum of ratings on administered items between the basal and ceiling, with all items credited successfully below the basal score. The raw score is then converted to a norm-referenced v-scale score (with a mean of 15 and standard deviation of 3). The subdomain scores, then, combine for the domain standard scores and overall ABC (with a mean of 100 and standard deviation of 15). The 3rd Edition introduced growth scale values (GSVs) as a new scoring type. GSVs provide a person ability score. GSVs are not standardized but instead represent abilities and behaviors on an interval scale, which is useful for longitudinal studies of within-person change (Farmer et al., 2020). Person ability scores hold great potential for clinical trials within genetic conditions associated with neurodevelopment (GCAND), including within SCN2A (Farmer et al., 2023).

Study procedures

The SCN2A-CTRS was co-designed with the FamilieSCN2A Foundation. It involved cross-sectional recruitment and longitudinal follow-up. The overarching goal was to collect validity evidence for a series of COAs that is required to demonstrate that an instrument is fit-for-purpose for a clinical trial. Families of children with an SCN2A-related disorder were recruited through the FamilieSCN2Aa outreach efforts. Parents reported information for their SCN2A-affected children at study entry and again at approximately 6 and 12 months after study entry. In addition to COAs, parents completed at baseline a medical history form. At each study event, they also completed a functional abilities form that included questions derived directly or in modified form from the rehabilitation literature for gross motor function, communication, eating, and hand use (Paulson and Vargus-Adams, 2017). The functional abilities form also included CDC developmental checklist items through about age five and some additional information about speech, language, and communication.

The Vineland-3 comprehensive interview was administered by trained research assistants, many of whom were actively pursuing advanced degrees in clinical psychology. All interviewers were first trained by a licensed clinical psychologist with extensive experience in developmental assessments as used in clinical and research settings. Practice interviews were recorded and reviewed by the licensed psychologist until interviewers were performing adequately. All interviews during the course of the study were audio-recorded, and 20% were rescored by the licensed psychologist to evaluate inter-rater reliability. Throughout the study, interviews met with the supervising psychologist and the study PI to review any questions with scoring and difficulties with administration to minimize drift in administration and scoring criteria as well as other potential errors. Forty-six families completed the Vineland twice, allowing estimation of the test-retest reliability of the GSV scores.

Statistical analyses

Because GSV scores are only available for the 11 subdomains, our analyses focused on these scores, although standardized scores for the composite and four primary domains are provided for context. Both GSVs and subdomain norm-referenced v-scale scores were evaluated for the floor and ceiling effects. We defined floor effects as obtaining the minimum possible score on that subdomain (i.e., a GSV of 10, or the lowest v-scale score possible at a given chronological age [often but not exclusively a v-scale of 1]).

We calculated internal consistency and floor effects (and ceiling effects should they occur) at the first assessment condition. Floor and ceiling effects are provided descriptively, but following Terwee et al. (Terwee et al., 2007), we consider these problematic if more than 15% of the sample obtain a score at the extreme of the test.

Second, we evaluated the internal consistency reliability using permutation-based random split half reliability (Parsons et al., 2019). The data were randomly split in half, and the two halves were correlated using the Pearson product-moment correlation coefficient, with the Spearman-Brown correction applied to adjust for the split test length. This process was repeated for 1000 permutations of random splits within each domain, thereby providing a sampling distribution for the internal consistency estimates. Split-half reliability can be interpreted as good when greater than 0.80 and excellent above 0.90.

Next, we calculated test-retest reliability at the short-term retest occasion. The study design opened the retest window at 14 days, with interviews scheduled based on family availability thereafter.

For quantifying test-retest reliability, it is generally expected to be a shorter window so that changes are unlikely to occur; given that this study was not providing active intervention, we allowed the retest assessment to occur up to 6 weeks after assessment. This is broadly consistent with the Vineland-3 manual, which reports test-retest reliability with the second assessment occurring 12 to 35 days after the initial assessment. We use the intra-class correlation coefficient (ICC) for absolute agreement for this purpose.

Finally, we calculated inter-rater reliability using Krippendorff's alpha (Hayes and Krippendorff, 2007) between the 20% of cases that were recoded by the supervising psychologist. Krippendorff's alpha is a reliability coefficient that subtracts the proportion of observed disagreement over the expected disagreement from 1 (i.e., perfect agreement). The observed and expected disagreement functions vary by the level of measurement, making it appropriate for all scale types, including ordinal data (like the item- and scale-level coded responses utilized herein). Asymptotically, Krippendorff's alpha encompasses other common statistics, like Scott's pi or Pearson's intraclass correlation coefficient (Hayes and Krippendorff, 2007), so interpretive guidance on those statistics is equally applicable to alpha (e.g., values greater than 0.75 indicate good reliability and above 0.90 as excellent; Koo & Li, 2016).

Results

Floor and ceiling effects

All results appear in Table 2. Ceiling effects were not present for any of the 11 subdomains of the Vineland and therefore do not appear in the table. However, floor effects were observed on all domains with either the GSV or norm-referenced v-scale score. Significant floor effects for the GSVs were found for written (from the communication domain), personal, domestic, and community (from the daily living skills domain), and fine motor (from the motor domain). Floor effects on the v-scale score were above the threshold for all subdomains, with the lowest rates (only 19%) observed for coping (from the socialization domain).

Internal consistency

Split-half reliability was high for all domains. As shown in Table 2, the median split-half reliability coefficient was above 0.90 for all subdomains excluding coping. The permuted 95% confidence intervals were also all above 0.60 (indeed, even above 0.70), which is often considered the lower bound for a measure to be considered reliable.

Test-retest reliability

Although we targeted a retest window of 14 days, the median follow-up window was 21 days (IQR: 15–29.75 days). All 11 subdomains exhibited a high level of reliability, as shown in Table 2. The ICCs provide an index of absolute agreement, but also relevant to discussions of test-retest reliability is expected score changes during a short-term period of stable interventions. Expected score changes over the course of 2 weeks were variable and small (all ΔGSV <2), well within the standard error of measurement for the subdomains.

Test-retest reliability can also be represented graphically using a Bland-Altman plot. The mean score across assessments

Table 2. Psychometric Characteristics of the Vineland-3

Domain	*Subdomain*	*GSV floor effects,* N *(%)*	*v-scale floor effects,* N *(%)*	*Internal consistency median (bootstrapped 95% CI)*	*Test-retest ICC*	*Inter-rater alpha overall*	*Inter-rater item-level item count, mean (SD)*
Communication	Expressive	1 (1.5%)	58 (89.2%)	0.98 (0.96, 0.99)	0.97	0.98	nI = 22 0.96 (0.08)
	Receptive	3 (4.6%)	52 (80.0%)	0.98 (0.97, 0.99)	0.97	0.96	nI = 24 0.96 (0.06)
	Written	20 (36.4%)	40 (72.7%)	0.97 (0.93, 0.98)	0.96	0.98	nI = 7 1.00 (0.01)
Socialization	Interpersonal	1 (1.5%)	32 (49.2%)	0.95 (0.92, 0.97)	0.94	0.99	nI = 20 0.97 (0.05)
	Play	5 (7.7%)	37 (56.9%)	0.96 (0.93, 0.98)	0.91	0.97	nI = 19 0.93 (0.08)
	Coping	4 (6.3%)	12 (19%)	0.87 (0.79, 0.92)	0.65	0.98	nI = 15 0.95 (0.09)
Daily living skills	Personal	11 (16.9%)	56 (86%)	0.98 (0.95, 0.98)	0.95	0.99	nI = 19 0.97 (0.06)
	Domestic	44 (80.0%)	44 (80.0%)	0.98 (0.96, 0.99)	0.99	1.00	nI = 5 0.80 (0.45)
	Community	36 (65.5%)	40 (72.7%)	0.98 (0.96, 0.99)	0.92	0.87	nI = 5 0.83 (0.11)
Motor	Gross	5 (7.7%)	50 (80.6%)	0.99 (0.97, 0.99)	0.98	0.78	nI = 28 0.96 (0.08)
	Fine	12 (18.5%)	52 (81.3%)	0.97 (0.92, 0.98)	0.98	0.88	nI = 20 0.98 (0.07)

Floor effects and split-half reliability was calculated on the full sample (n = 65). Test-retest reliability was calculated on the subset n = 46 with two short-term follow-up assessments. Inter-rater reliability overall was calculated on double-coded cases (n = 49). We required a minimum of 10 individuals to have taken an item in order to evaluate item-level inter-rater reliability. As such, most items within a subdomain could not be evaluated. The number of available items is indicated within the row as appropriate.

is plotted on the x-axis, while the difference between scores is plotted on the y-axis. Ideally, the mean difference should be about zero, suggesting no change over time. The closer all plotted values are to zero, the greater the test-retest reliability. If a pattern is observed in the Bland-Altman plot, it would suggest that reliability was greater in one range of ability than in another, which may be masked by the overall reliability estimates. Bland-Altman plots for the 11 subdomains are provided as panels of Figure 1.

Inter-rater reliability

Approximately 20% of assessments within each research assistant (n = 49) were recoded by the supervising licensed psychologist. Krippendorff's alpha was calculated by item and by subdomain. Subdomain-level inter-rater reliability was exceedingly high (>0.95) for most domains, but community (within the daily living skills domain) and both gross and fine motor (within the motor domain) had lower reliability coefficients. These were still above 0.80 and acceptable for clinical research. Item-level reliability was also high but varied across the subdomains. The greatest variability across subdomains was observed on items within the domestic and community subdomains. However, we required a minimum of 10 individuals to have taken an item before we would evaluate item-level reliability. As such, only five items within these subdomains

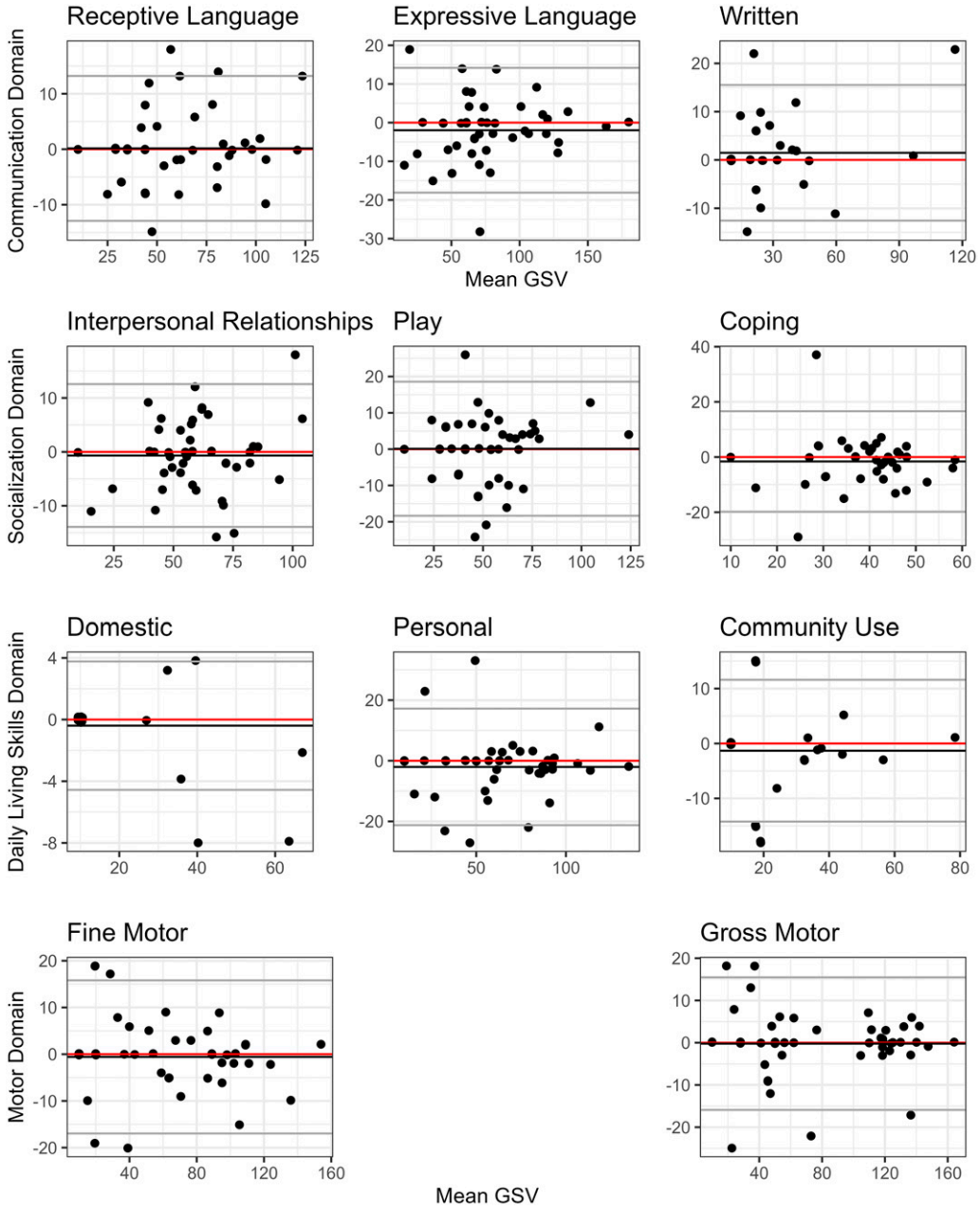

FIG. 1. Bland-Altman plots for the Vineland-3. Bland-Altman plots demonstrate the concordance in GSV scores across retest occasions. The mean of the two scores is on the x-axis, and the score difference is on the y-axis. The black solid line is the obtained mean difference (which indicates potential bias), and the red solid line indicates the ideal or unbiased difference. The solid gray lines indicate the 95% limits of agreement. GSV, growth scale value.

were eligible for evaluation. The number of items considered for item-level reliability is also reported within Table 2.

Discussion

SCN2A-RD is characterized by severe to profound global impairments. In our sample, the standardized Vineland-3 ABC score indicated function at <5 SD of the test normative mean, a range that is far outside the general population expectations. The substantial floor effects demonstrate the inappropriateness of the standardized scores for use in this target population. By contrast, GSVs had minimal floor effects on the domains that families anecdotally said were most relevant to them and which are consistent with previous publications on patient priorities (Downs et al., 2024). The domains that continued to exhibit floor effects even on the GSV scale—that is, written, domestic, and community—are also domains that have shown floor effects in other GCAND, such as creatine transport disorder syndrome (c.f., Fig. 1 from Farmer et al., 2020). Overall, and similar to the Vineland-II (c.f. Berg et al., 2021), the Vineland-3 subdomains exhibited high internal consistency, test-retest reliability, and inter-rater reliability, supporting their potential for use as COAs in future clinical trials.

The Vineland is one of the most commonly used measures of adaptive behavior within GCAND broadly. It has been used for eligibility screening, participant stratification, and baseline characterization in clinical trials. It is also being used within clinical trials. The FDA has clear guidelines about the level of evidence needed to consider a measure a COA, and this study provides some of that quantitative evidence.

The FDA emphasizes test-retest and inter-rater reliability, both of which were excellent for all subdomains excluding coping test-retest. Indeed, within this SCN2A sample, the reliability was higher than the reliability reported in the standardization sample for the Vineland-3 (e.g., median test-retest reliability across ages 0–12 years was 0.74 and median inter-rater reliability for ages 0–20 years was 0.76) (Sparrow et al., 2016). Subdomain-level inter-rater reliability was excellent but slightly lower for the two motor subdomains. We believe there are several causes for these findings. First, the sample had significant deficits overall, and therefore much of the scale was not represented in this sample. Second, significant credit for the high inter-rater reliability is credited to the excellent training and supervision provided by the supervising psychologist. Weekly conference calls and standardized training support consistent coding across cases. Future clinical trials seeking to use the Vineland should use the Comprehensive Interview Form as opposed to the parent-/caregiver-reported form, and the interviewers should undergo clear training on coding ambiguous items. This is particularly relevant in SCN2A (as well as other GCAND), where phenotype may modify the "typical" presentation of adaptive behaviors. Alternative communication devices are more common in these populations, and accurately crediting an individual's functioning requires more than just a cursory understanding of the item content. Training and fidelity checks will be key components for any future clinical trial.

Even though the FDA puts a lower emphasis on internal consistency and reliability, this is still an important psychometric attribute. We calculated the random split-half reliability across 1000 permutations for the item-level Vineland data. This was also very high, supporting the use of the Vineland in clinical trials.

However, floor effects were common across many subdomains, which may limit their applicability. Our primary interest was in the GSV scores, which are most appropriate for clinical trials; norm-referenced scores, though useful for other purposes, are less useful in the context of clinical trials (c.f. Farmer et al., 2023; Farmer et al., 2020) and may actually reduce the ability to detect an effect in clinical trials (Farmer et al., 2023; Kwok et al., 2022). Using the GSVs, written, domestic, and community were particularly problematic and well above the 15% threshold proposed within the literature (Terwee et al., 2007). Personal and fine motor were also above the proposed threshold, but to a lesser extent. Given the small sample size and the importance of these domains within sodium channel protein type 2 subunit alpha related disorders, we consider these results preliminary. More research is necessary as to whether these floor effects would be replicated.

Floor effects were even more common using the norm-referenced v-scale scores. In this study, we considered a floor score to be the age-specific minimum possible subdomain v-scale. This is noteworthy because even though some subdomains for some ages can obtain a v-scale of 1, using that definition would erroneously conclude that since an individual didn't receive a score of one, they were not on the floor of the subdomain. For some tests, the floor effects were similar between GSV and v-scale, but for others they were quite discrepant—especially for domains such as personal (daily living skills domain) or gross motor (motor domain). Among populations with significant impairments, including GCAND broadly and sodium channel protein type 2 subunit alpha related disorders specifically, it is common to see a discrepancy between personability scores and norm-referenced scores, which members of our team have previously shown would negatively impact statistical power in clinical trials (Farmer et al., 2023). Even in less-impaired populations, treatment effects may be reduced when using norm-referenced scores (Kwok et al., 2022), likely because of floor effects.

Given these findings, we strongly advise clinical trials to consider which score is most appropriate for the Vineland's context of use in their study. The FDA differentiates several contexts of use, including patient eligibility, enrichment, or stratification, and monitoring efficacy, among others. The v-scale score may be useful for patient selection, but it is not useful for monitoring outcomes or efficacy evaluations. This study provides further support that to use the Vineland and a COA or as part of a (likely secondary) endpoint, the GSV must be the scoring metric.

Limitations

This study is not without its limitations. The primary one is the small sample size for traditional psychometric analyses. The overall sample of 65 individuals with sodium channel protein type 2 subunit alpha related disorders is large for such a rare condition. Sample sizes were smaller for some analyses, with test-retest reliability calculated on 46 (70.8% of the sample) and inter-rater reliability calculated on 49 assessments (20% of assessments, with some individuals contributing more than one assessment). This sample size is consistent with or greater than other observational studies conducted within SCN2A-RDs, but a larger sample would be necessary to use more modern psychometric methods, such as structural validity assessments or item response theory modeling. Future research should consider newer methods that can iteratively include longitudinal data in modern psychometric analyses (Houts et al., 2018).

Additionally, this study focused on the quantitative aspects of a COA, but as part of the FDA's guidance on patient-focused drug development, there is also a need to demonstrate the relevance and importance of potential concepts of interest to patients, families,

and other stakeholders (FDA, 2022a). This is usually accomplished using qualitative or mixed-method research. Anecdotally, families and caregivers of participants in the study mentioned domains such as writing and domestic as less-relevant to SCN2A—there are so many other important functional skills that the families emphasized instead. Families did endorse communication and behavior as highly significant, which the Vineland measures, but not all domains are equally important. Formal evaluation of patient's and family's priorities is an important next step for establishing clinical trial readiness and an appropriate context of use for the Vineland. Further, qualitative and quantitative methods are necessary to anchor change on important domains to clinical meaningfulness. Patient and caregiver impressions of change should anchor evaluations on whether score changes are meaningful. This is an important next step for future research.

Conclusions

This was a comparatively large study, exclusively within SCN2A-RDs, which is characterized by severe to profound functional impairments. The results provide strong evidence for the psychometric properties of the Vineland in individuals with sodium channel protein type 2 subunit alpha related disorders. The FDA emphasis on functioning as part of its patient-focused drug development, and adaptive behavior is a key type of functional outcome. The FDA also places a high emphasis on psychometric evidence within the target population (as opposed to relying on the general population or measure development sample; FDA, 2023), in particular test-retest and inter-rater reliability. Our findings demonstrate the unsuitability of norm-referenced scores (primarily due to floor effects) but provide strong evidence for the GSVs as being fit-for-purpose in this population. We recommend that the Vineland-3 GSVs be utilized both in research and in clinical care, as this can provide real-world data useful for quantifying adaptive behavior in SCN2A-RDs.

Clinical Significance

Disease-modifying treatments are in development for monogenic epilepsies such as sodium channel protein type 2 subunit alpha related disorders. However, clinical trials need appropriate outcome measures to assess efficacy across a wide range of development. This result of this study supports the use of the Vineland-3 Comprehensive Interview as a potential clinical outcome assessment in ongoing and future trials within SCN2A.

Ethics Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Ethical review and approval were granted by the North Star Ethics Review Board, protocol NB200048, for the SCN2A Clinical Trials Readiness study.

Authors' Contributions

A.J.K.: Formal analysis, writing—original draft; A.N.N.: Methodology, investigation, writing—review and editing; K.P.: Methodology, investigation, writing—review and editing; A.J.E.K.: Methodology, investigation, writing—review and editing; E.A.: Validation, investigation, writing—review and editing; L.S.M.: Conceptualization, funding acquisition, writing—review and editing; A.T.B.: Conceptualization, funding acquisition, methodology, data curation, writing—review and editing.

Disclosures

This study was funded by the FamilieSCN2A Foundation. A.J.K., L.E., A.N.N., K.P., A.J.E.K., and E.A. served as consultants to the FamilieSCN2A Foundation for the purposes of this project. L.S.M. and A.T.B. are employed by the FamilieSCN2A Foundation, serving as the Founder and Executive Director and as the Director of Clinical Outcomes Measures, respectively. All authors approved the article and consent to publish was not required from the Foundation.

References

ADMINISTRATION, U. S. FOOD AND DRUG (FDA). Guidance 1: Patient-Focused Drug Development: Collecting Comprehensive and Representative Input (final guidance). *US* FDA: *Washington, DC*; 2018.

ADMINISTRATION, U. S. FOOD AND DRUG (FDA). Guidance 3: Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments (draft guidance). *US* FDA: *Washington, DC*; 2022b.

ADMINISTRATION, U. S. FOOD AND DRUG (FDA). Guidance 4: Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints for Regulatory Decision-Making (draft guidance). *US* FDA: *Washington, DC*; 2023.

ADMINISTRATION, U. S. FOOD AND DRUG (FDA). Rare Diseases: Considerations for the Development of Drugs and Biological Products (final guidance). *US* FDA: *Washington, DC*; 2023.

ADMINISTRATION, U. S. FOOD AND DRUG (FDA). Rare Diseases: Natural History Studies for Drug Development (draft guidance for industry). *US* FDA; 2019.

ADMINISTRATION, U. S. FOOD AND DRUG (FDA). Guidance 2: Patient-Focused Drug Development: Methods to Identify What Is Important to Patients (final guidance). *US* FDA: *Washington, DC*; 2022a.

American Psychiatric Association (APA). Diagnostic and Statistical Manual (DSM-5). 5th ed. American Psychiatric Association: Washington; 2013.

Basch E, Geoghegan C, Coons S, et al. Patient-reported outcomes in cancer drug development and us regulatory review: Perspectives from industry, the food and drug administration, and the patient. JAMA Oncology, 2015;1(3):375–379.

Berg AT, Palac H, Wilkening G, et al. SCN2A-Developmental and Epileptic Encephalopathies: Challenges to trial-readiness for non-seizure outcomes. Epilepsia 2021;62(1):258–268.

Devinsky O, Cross JH, Laux L, et al; Cannabidiol in Dravet Syndrome Study Group. Trial of cannabidiol for drug-resistant seizures in the Dravet syndrome. N Engl J Med 2017;376(21):2011–2020.

Devinsky O, Patel AD, Cross JH, et al; GWPCARE3 Study Group. Effect of cannabidiol on drop seizures in the Lennox-Gastaut syndrome. N Engl J Med 2018;378(20):1888–1897.

Downs J, Ludwig NN, Wojnaroski M, et al. What does better look like in individuals with severe neurodevelopmental impairments? A qualitative descriptive study on SCN2A-related developmental and epileptic encephalopathy. Qual Life Res 2024;33(2):519–528.

Farmer CA, Kaat AJ, Thurm A, et al. Person ability scores as an alternative to norm-referenced scores as outcome measures in studies of neurodevelopmental disorders. Am J Intellect Dev Disabil 2020; 125(6):475–480.

Farmer C, Thurm A, Troy JD, et al. Comparing ability and norm-referenced scores as clinical trial outcomes for neurodevelopmental disabilities: A simulation study. J Neurodev Disord 2023;15(1):4–9.

Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 2007;1(1):77–89.

Heron SE, Crossland KM, Andermann E, et al. Sodium-channel defects in benign familial neonatal-infantile seizures. Lancet 2002; 360(9336):851–852.

Hidecker MJ, Paneth N, Rosenbaum PL, et al. Developing and validating the communication function classification system for individuals with cerebral palsy. Dev Med Child Neurol 2011;53(8): 704–710.

Houts CR, Morlock R, Blum SI, et al. Scale development with small samples: A new application of longitudinal item response theory. Qual Life Res 2018;27(7):1721–1734.

Hudson KL, Collins FS. The 21st century cures act—a view from the NIH. N Engl J Med 2017;376(2):111–113.

Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 2016;15(2):155–163.

Kwok E, Feiner H, Grauzer J, et al. 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 2022;65(11):4268–4279.

O'Callaghan FJK, Lux AL, Darke K, et al. The effect of lead time to treatment and of age of onset on developmental outcome at 4 years in infantile spasms: Evidence from the United Kingdom Infantile Spasms Study. Epilepsia 2011;52(7):1359–1364.

Parsons S, Kruijt A-W, Fox E. Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. Advances in Methods and Practices in Psychological Science 2019;2(4):378–395.

Paulson A, Vargus-Adams J. Overview of four functional classification systems commonly used in cerebral palsy. Children (Basel), 2017;4(4).

Perfetto EM, Burke L, Oehrlein EM, et al. Patient-focused drug development: A new direction for collaboration. Med Care 2015;53(1):9–17.

Sanders SJ, Campbell AJ, Cottrell JR, et al. Progress in understanding and treating SCN2A-mediated disorders. Trends Neurosci 2018; 41(7):442–456.

Schalock RL, Borthwick-Duffy SA, Bradley VJ, et al. 2010. Intellectual disability: Definition, classification, and systems of supports. ERIC.

Semmel ES, Fox ME, NA SD, et al. Caregiver-and clinician-reported adaptive functioning in Rett syndrome: A systematic review and evaluation of measurement strategies. Neuropsychol Rev 2019; 29(4):465–483.

Sparrow S, Cicchetti D, Saulnier CA. Vineland Adaptive Behavior Scales, Third Edition. Pearson; 2016.

Symonds JD, Zuberi SM, Stewart K, et al. Incidence and phenotypes of childhood-onset genetic epilepsies: A prospective population-based national cohort. Brain 2019;142(8):2303–2318.

Tassé MJ, Schalock RL, Balboni G, et al. The construct of adaptive behavior: Its conceptualization, measurement, and use in the field of intellectual disability. Am J Intellect Dev Disabil 2012;117(4): 291–303.

Terwee CB, Bot SDM, DE Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60(1):34–42.

Towns M, Rosenbaum P, Palisano R, et al. Should the Gross Motor Function Classification System be used for children who do not have cerebral palsy? Develop Med Child Neuro 2018;60(2):147–154.

Wolff M, Johannesen KM, Hedrich UBS, et al. Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A. Brain 2017;140(5):1316–1336.

Yang S, Paynter JM, Gilmore L. Vineland adaptive behavior scales: II profile of young children with autism spectrum disorder. J Autism Dev Disord 2016;46(1):64–73.

Address correspondence to:
Aaron J. Kaat, PhD
Department of Medical Social Sciences
Northwestern University Feinberg School of Medicine
625 N. Michigan Ave.
Suite 2700
Chicago, IL 60611
USA

E-mail: aaron.kaat@northwestern.edu