DOI 10.1007/s11689-008-9001-8

A solution to limitations of cognitive testing in children with intellectual disabilities: the case of fragile X syndrome

David Hessl · Danh V. Nguyen · Cherie Green · Alyssa Chavez · Flora Tassone · Randi J. Hagerman · Damla Senturk · Andrea Schneider · Amy Lightbody · Allan L. Reiss · Scott Hall

Received: 12 August 2008 / Accepted: 27 October 2008 / Published online: 22 November 2008

Abstract Intelligence testing in children with intellectual disabilities (ID) has significant limitations. The normative samples of widely used intelligence tests, such as the Wechsler Intelligence Scales, rarely include an adequate number of subjects with ID needed to provide sensitive measurement in the very low ability range, and they are highly subject to floor effects. The IQ measurement problems in these children prevent characterization of strengths and weaknesses, poorer estimates of cognitive abilities in research applications, and in clinical settings, limited utility for assessment, prognosis estimation, and planning intervention. Here, we examined the sensitivity of the Wechsler Intelligence Scale for Children (WISC-III) in a large sample of children with fragile X syndrome (FXS), the most common cause of inherited ID. The WISC-III was administered to 217 children with FXS (age 6–17 years, 83 girls and 134 boys). Using raw norms data obtained with permission from the Psychological Corporation, we calculated normalized scores representing each participant's actual deviation from the standardization sample using a *z*-score transformation. To validate this approach, we compared correlations between the new normalized scores versus the usual standard scores with a measure of adaptive behavior (Vineland Adaptive Behavior Scales) and with a genetic measure specific to FXS (*FMR1* protein or FMRP). The distribution of WISC-III standard scores showed significant skewing with floor effects in a high proportion of participants, especially males (64.9%–94.0% across subtests). With the z-score normalization, the flooring problems were eliminated and scores were normally distributed. Furthermore, we found correlations between cognitive performance and adaptive behavior, and between cognition and FMRP that were very much improved when using these normalized scores in contrast to the usual standardized scores. The results of this study show that meaningful variation in intellectual ability in children with FXS, and probably other populations of children with neurodevelopmental disorders, is obscured by the usual translation of raw scores into standardized scores. A method of raw score

D. Hessl (✉) · C. Green · A. Chavez · F. Tassone · R. J. Hagerman · A. Schneider
Medical Investigation of Neurodevelopmental Disorders (M.I.N.D.) Institute, University of California-Davis Medical Center,
2825 50th Street,
Sacramento, CA 95817, USA
e-mail: david.hessl@ucdmc.ucdavis.edu

D. Hessl
Department of Psychiatry and Behavioral Sciences,
University of California Davis Medical Center,
Sacramento, CA, USA

R. J. Hagerman
Department of Pediatrics,
University of California Davis Medical Center,
Sacramento, CA, USA

F. Tassone
Department of Biochemistry and Molecular Medicine,
University of California Davis School of Medicine,
Davis, CA, USA

D. V. Nguyen
Public Health Sciences, Division of Biostatistics,
University of California Davis,
Davis, CA, USA

D. Senturk
Department of Statistics, Pennsylvania State University,
State College, PA, USA

A. Lightbody · A. L. Reiss · S. Hall
Center for Interdisciplinary Brain Sciences Research
and Department of Psychiatry and Behavioral Sciences,
Stanford University School of Medicine,
Stanford, CA, USA

transformation may improve the characterization of cognitive functioning in ID populations, especially for research applications.

Keywords *FMR1* gene · IQ · Mental retardation · Assessment · FMRP

Introduction

The accurate measurement of cognitive capacity in children with intellectual disabilities (ID) is important for determining appropriate diagnosis, service eligibility, individual strengths and weaknesses, treatment and education planning, and for research studies on these populations that rely heavily on IQ as a critical variable of interest. ID is a disability, originating before the age of 18, characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills (American Association of Intellectual and Developmental Disabilities; www.aaidd.org). The Diagnostic and Statistical Manual of Mental Disorders (DSM-IV; American Psychiatric Association [1]) classifies ID in the following degrees of severity based on adaptive functioning and IQ: Mild (50–55 to approximately 70; ~85% of the ID population), Moderate (35–40 to 50–55; ~10% of ID), Severe (20–25 to 35–40; 3%–4% of ID), and Profound (below 20 or 25; 1%–2% of ID) [1]. Intellectual functioning is defined as IQ obtained by assessment with a standardized, individually administered intelligence test such as the Wechsler Intelligence Scales, the Stanford–Binet, or the Kaufman Assessment Battery. Although the DSM-IV includes classifications for more impaired individuals, it is very challenging to measure the IQ reliably and accurately in subjects with ID below the Mild range (IQ 50–70). Indeed, a major limitation of these tests is that they do not typically measure IQ below 40 or 50, and that subtest standardized scores, which contribute to the overall score, are highly subject to floor effects and poor estimates of true ability.

A further complication and limitation is that whereas IQ tests generally do not measure functioning below 4 standard deviations below average (IQ=40), measures of adaptive behavior, such as the Vineland Adaptive Behavior Scales (VABS) [2], typically have a standard score floor of 20 (over 5 standard deviations below average), making comparisons between cognitive capacity and daily functioning impossible for these individuals. The lack of sensitivity of intelligence tests in this range of functioning is typically due to relative dearth of children with ID of varying levels of severity in the standardization samples, and limitations in the range of difficulty of test items and tasks that prevent measurement of lower levels of ability. Notably, test publishers have recently made some improvements in the normative sampling of lower functioning children (Stanford–Binet, Fifth Edition [3]; Differential Ability Scales, Second Edition (DAS-II; [4]), and one of these tests now has a lower IQ limit of 30 (DAS-II).

Clinical and research experience with intelligence testing in children with neurodevelopmental disorders shows that meaningful variation in performance is often obscured by flooring effects when raw scores are converted to standardized scores based on the normative data in test manuals. We can use the performance of two 15-year-old children with ID on the Wechsler Intelligence Scale for Children, Third Edition (WISC-III) and the VABS to illustrate this point (on both of these measures, IQ and VABS standardized scores have a mean of 100 and standard deviation of 15. On the WISC-III, subtest standardized scores have a mean of 10 and standard deviation of 3, with a range of 1 to 19). “Sam” is 15 years of age, speaks in one- to two-word utterances, receives a VABS Adaptive Behavior Composite (ABC) score of less than 20 (below the 0.1 percentile) and a Full Scale IQ (FSIQ) of 40 (the floor of the test). On the WISC-III Vocabulary subtest, for example, he obtains a raw score of 1 which converts to a standardized score of 1 (in response to “What is a clock?” he answers, “Time.”, and then has no further correct responses). “Joe” is a verbally fluent 15-year-old with a VABS ABC score of 60. He obtains a Vocabulary raw score of 16 and responds to questions with complex phrases or complete sentences; however his raw score also converts to a standardized score of 1, the same as Sam. Joe obtains a FSIQ of 42, just 2 points higher than Sam.

Floor effects and other measurement problems in intelligence testing with children with ID are common; however with a few exceptions such as those below, they are not often recognized or discussed in published studies. In a longitudinal study of a large sample of adults with mental retardation using the Wechsler Adult Intelligence Scale—Revised (WAIS-R), Facon [5] reported mean IQ scores between 54 and 58 for four different age bands; however the scores and distributions were indicative of significant flooring effects that the authors acknowledged as a limitation in their discussion. In their analysis, the authors chose to use subtest raw scores instead of the standardized scores; they re-standardized the raw scores relative to their entire sample and summed these scores to create new composite verbal and performance scores for each subject. Another example comes from a study of 195 individuals with Down syndrome that were longitudinally assessed with the Stanford Binet, Fourth Edition. The authors reported that 37% of the available test results were assigned the lowest possible score of 36 [6] but that these individuals demonstrated highly variable levels of performance despite flat standardized score profiles.

Our research centers have been studying individuals with FXS, the leading cause of inherited ID, for the past 25 years. FXS is a single gene disorder caused by a mutation in the fragile X mental retardation 1 (*FMR1*) gene on the X chromosome at Xq27.3. This mutation results from a trinucleotide expansion preventing normal transcription, and leads to reduction or absence of the *FMR1* protein (FMRP) [7, 8] and consequent abnormal brain development, including aberrant dendritic arborization and synaptic plasticity [9–13]. In full mutation females, FMRP is usually expressed only by the normal allele carried on the active X chromosome. As a result, females tend to be higher functioning than males with FXS, although there is wide variability from significant ID to normal or above average IQ. Variable FMRP expression also results from mosaicism, where transcriptional silencing of the gene does not occur in all cells, either because of varying sizes of the repeat expansion or variation in methylation. Although more frequent in males, mosaicism also occurs in females with FXS. Individual differences in FMRP production in the brain as a result of these factors are thought to account for a significant proportion of the variability in IQ in individuals with FXS.

We have sought to understand the impact of gene function, brain function, and environmental variation on cognition and behavior in FXS, with the ultimate goal of identifying effective interventions based on this information. However, our research and clinical work has been significantly limited by a lack of IQ measurement sensitivity, as described above, in a substantial portion of individuals with this disorder. For example, in one study, designed to determine genetic and environmental factors contributing to IQ (as measured by the Wechsler scales), 43% of boys with FXS scored at the floor on all 12 subtests, and all of these children obtained a FSIQ of 40 [14–16]. Although these individuals demonstrated considerable variability in their cognitive abilities and level of adaptive behavior [15], their individual strengths and weaknesses and variation within the group were not reflected in their standardized scores. In an attempt to overcome this problem, in a recent study [17] we abandoned standard scores altogether, and employed raw WISC-III subtest scores to examine the development of intellectual functioning in children with FXS. Using raw scores, and covarying for age, we found that intellectual functioning in children with FXS developed approximately two times slower than typically developing siblings over the age range of 6 to 16 years. While raw scores may offer significant advantages over standard scores (e.g., no floor effect, normal distribution of scores), the WISC-III manual does not contain raw subtest scores from the normative population. Thus, investigators cannot use raw subtest scores in their analyses without the inclusion of a well-matched comparison group.

Fragile X offers a unique opportunity to examine the sensitivity of intelligence testing in an ID population. The specific genetic etiology has been identified, the neuroanatomical morphology has been well-described, and the cognitive and behavioral phenotype is well known and relatively consistent. Although there are differences in FMRP expression in the brain compared to blood, the gene-dose of the mutation can be estimated by measurement of FMRP in lymphocytes. The degree of FMRP deficit can then be correlated with the cognitive deficit as measured by standardized testing [18, 19]. Thus, FXS is a model for examining assumptions about measurement of cognition of individuals with mental impairment that can then be tested in other neurodevelopmental disorders (e.g. autism, Down syndrome) and more heterogeneous populations (e.g. children with idiopathic ID).

Here, we examined the sensitivity of the WISC-III, one of the most widely used intelligence tests, in a large sample of children and adolescents with FXS. First, we show the distribution of the usual standard scores in this sample of boys and girls. Next, we present a method for calculating new normalized scores representing each child's actual deviation from the standardization sample, based on the raw score descriptive statistics obtained with permission from the publisher of the WISC-III (Psychological Corporation, San Antonio, TX). Finally, we compare the distribution of the normalized scores to the usual standardized scores, and correlate each of these with another measure of developmental level, the Vineland Adaptive Behavior Scales, and the degree of FMRP deficit.

Methods

Participants

Participants included 217 children with the fragile X full mutation ranging in age from 6 to 17 years (83 girls, mean age=10.94±3.01 years; 134 boys, mean age 11.04±2.59 years). Twelve girls (14.5%) and 44 boys (32.8%) had repeat size mosaicism and 1 girl (1.2%) and 13 boys (9.7%) had methylation mosaicism. Sixty-nine girls and 103 boys participated in studies conducted at the Center for Interdisciplinary Brain Sciences Research at Stanford University (PI, A. Reiss) and 14 girls and 31 boys in studies at the M.I.N.D. Institute at University of California Davis (PIs R. Hagerman and D. Hessl). Note that participants resided in various locations throughout the United States and Canada and were from a wide range of socioeconomic backgrounds as previously described [14–16]. The ethnic distribution was 86.6% Caucasian, 5.1% Hispanic, 1.9% African American, 0.9% Asian, 0.5% Native American, and 5.1% other or unknown. The mothers' highest level of education obtained was 33.3% college degree, 32.8% partial college or specialized training, 17.7% high school degree or GED, 14.6% graduate professional training, and 1.6% partial high school. FSIQ ranged from 40 to 123 (mean 50.0, SD 19.5). The parents of all participants provided written consent, and participants provided assent when possible, according to protocols approved by Institutional Review Boards at Stanford University and U.C. Davis.

Measures

Wechsler Intelligence Scale for Children, Third Edition (WISC-III; [20]) The WISC-III is a standardized test of intellectual aptitude for children between ages 6 and 16 years, 11 months. It is an individually administered clinical instrument with 13 subtests (all but the optional Mazes subtest were used in the study), each of which assesses either Verbal or Performance (perceptual-motor) abilities. A description of abilities addressed by each subtest is shown in Table 1. Each subtest generates a raw score, which then yields a standardized score based on normative data, and these standardized scores are combined and translated into overall Verbal IQ (VIQ), Performance IQ (PIQ) and Full Scale (FSIQ) scores. The WISC-III standardization sample included 2200 individuals, including 200 children in each of 11 age groups between ages 6 and 16 years. The groups were stratified by sex, race/ethnicity, geographic region, and parent education based on the 1988 U.S. Bureau of the Census. In the standardization sample, 7% were classified as learning disabled, speech/language impaired, emotionally disturbed or physically impaired. Published materials do not include information on any children with ID in the sample.

Vineland Adaptive Behavior Scales (VABS; [2]) The VABS is a widely used tool for assessing an individual's ability to care for one's self personally and socially. The VABS was designed to be administered as a semi-structured informant interview for assessing strengths and weaknesses of individuals from birth through 18 years 11 months or low-functioning adults. Part of the utility of this measure is the ability to gain accurate reporting from a responder who is familiar with a person's behavior. The interview lasts between approximately 60 min and contains 297 items. Adaptive behavior is measured in four to five domains: Communication (receptive, expressive and written), Daily Living Skills (personal, domestic, and community), Socialization (interpersonal relationships, play and leisure time, and coping skills), and Motor Skills (gross motor and fine motor; completed only for the youngest children). An Adaptive Behavior Composite (ABC) is yielded by combining scores on each of the four (or five) main domains. Standardization samples of handicapped and non-handicapped individuals provided normative data for the VABS and included 3,000 individuals between birth and 18 years 11 months, stratified by sex, race or ethnic group, community size, geographical region, and parents' education level.

Table 1 WISC-III subtests and abilities measured

Subtest	Abilities
Verbal	
Information	Range of knowledge, long term memory
Similarities	Abstract reasoning, concept formation
Arithmetic	Numerical reasoning and computation
Vocabulary	Word knowledge
Comprehension	Practical knowledge, social judgment
Digit Span	Auditory short-term memory
Performance	
Picture Completion	Visual perception, attention to detail
Coding	Visual–motor information processing
Picture Arrangement	Nonverbal reasoning and sequencing
Block Design	Spatial visualization and reasoning
Object Assembly	Visual perception and organization
Symbol Search	Scanning, matching, attention to detail

Fragile X diagnosis and FMRP analysis Southern blot analyses were performed according to procedures described by Taylor and colleagues [21]. FMRP expression from peripheral blood was determined by immunocytochemistry as the percent of FMRP-positive lymphocytes [22–24].

Statistical methods

Normalized scores The normalized scores were obtained using an age-dependent (within each population age band) *z*-score transformation as follows. Descriptive statistics (means and standard deviations) of subtest raw scores for each age band (6 years, 0 months to 6 years, 3 months; 6 years, 4 months to 6 years, 7 months, etc.) from the WISC-III standardization sample [20] were obtained with written permission from the Psychological Corporation (San Antonio, TX) for the purposes of this study. (Standardization data from the Wechsler Intelligence Scale for Children - Third Edition. Copyright © 1990 by Harcourt Assessment, Inc. Used with permission. All rights reserved). Denote the mean and standard deviation of a specific WISC-III subtest raw score in the *j*th age band by μ_j and σ_j, respectively. The normalized score for individual *i* falling into the *j*th age band is $z_{ij}=(r_{ij}-\mu_j)/\sigma_j$, where r_{ij} is the subtest raw score (note that data for each population age band contain representation of sex, race/ethnicity, education level and geographic region).

For example, a 12 year, 1 month old child obtains a Block Design subtest raw score of 3. In the standardization sample, for children 12 years, 0 months to 12 years, 3 months, the Block Design mean raw score is 39.86 and the standard deviation is 9.74. Therefore, the child's Block Design normalized *z* score is (3–39.86)/9.74=−3.78, or 3.78 standard deviations below the mean for his age-peers in the WISC-III normative sample.

Each participant's mean normalized scores are also calculated to indicate the overall deviation from the normative sample across subtests, analogous to the subtest standardized score combinations used to generate the VIQ, PIQ and FSIQ.

Analysis Summary and graphical analyses were used to characterize the raw, standardized and normalized scores. Downstream bivariate association/correlation analysis between Vineland ABC score and FMRP with normalized subtest scores were based on Pearson correlation as each variable is quantitative and continuous. We considered these correlative analyses to be descriptive, hence no formal *p*-value adjustment was used, although the majority of *p*-values based on normalized scores remained significant after false discovery rate (FDR) adjustment [25]. Finally, we compared the distribution of intellectual ability classifications (Mild ID, Moderate ID, Borderline, Low Average, etc.) determined by FSIQ in comparison to the assumed classification determined by the mean normalized score.

Results

Flooring effect of standardized scores

As expected, examination of the subtest standardized and IQ scores demonstrated significant flooring effects. The effects of flooring of subtest raw scores, resulting from standardization (i.e. the use of standard scores) are summarized in Table 2. As can be seen, a wide range of raw scores for each subtest received a floored value of 1 as the standard score, resulting in a loss of information on low performance, a range of cognitive abilities of interest in FXS, and potentially other ID populations of interest. The percent of participants in the study with floored standard scores ranged from 40.1% (Picture Completion) to 70.0% (Arithmetic). Although the bulk (e.g. 75th percentile) of floored raw scores were low (e.g. typically 0–7), the range of raw scores floored (Table 2, second column) were quite wide. Thus, important variability in the measurement of ability on subtests in lower functioning individuals was lost by the standard score flooring. As was expected for FXS individuals, a greater proportion of raw scores for males were floored compared to females: for example, 94.0% compared to 31.3% for Arithmetic and 84.5% compared 22.4% for Comprehension. Typically, the proportion of floored raw scores for males was many-fold higher than for females (see Table 2, column 1).

Characteristics and interpretation of normalized scores

The distribution of normalized scores for males and females are displayed in Fig. 1 for two representative subtests, Arithmetic and Vocabulary, alongside the standardized scores. Note that this flooring characteristic of standard scores was apparent for all subtests, as the distributions were non-normally distributed with significant positive skewing (See supplementary materials for all figures at http://dnguyen.ucdavis.edu/.html/SUP_iq/SupplementalFigures.pdf). In contrast, the normalized scores exhibited more "normal" distributions with no flooring effects. The flooring characteristics from raw to standard scores for both males and females induced skewed distributions of standard

Table 2 Flooring of subtest raw scores from standardization

Subtest domain	Percent of standard score of 1 (floored) overall % (males, females)	Raw score range floored[a]	75th percentile raw score range floored[b]
Arithmetic	70.0 (94.0, 31.3)	0–12	0–4
Block design	61.3 (82.1, 27.1)	0–14	0–5
Coding	46.7 (69.0, 11.1)	0–37	0–14
Comprehension	61.5 (84.5, 22.4)	0–15	0–8
Digit span	51.5 (73.4, 14.9)	0–7	0–3
Information	43.8 (64.9, 9.6)	0–10	0–6
Object assembly	47.4 (66.2, 17.1)	0–18	0–8
Picture arrangement	48.4 (70.2, 13.3)	0–10	0–4
Picture completion	40.1 (64.9, 15.7)	0–13	0–7
Similarities	49.5 (70.9, 14.6)	0–9	0–5
Symbol search	53.2 (76.5, 16.4)	0–11	0–4
Vocabulary	51.6 (74.6, 14.5)	0–21	0–12

[a] Corresponding range of raw scores with standard score of 1

[b] 75th Percentile of range of raw scores with standard score of 1

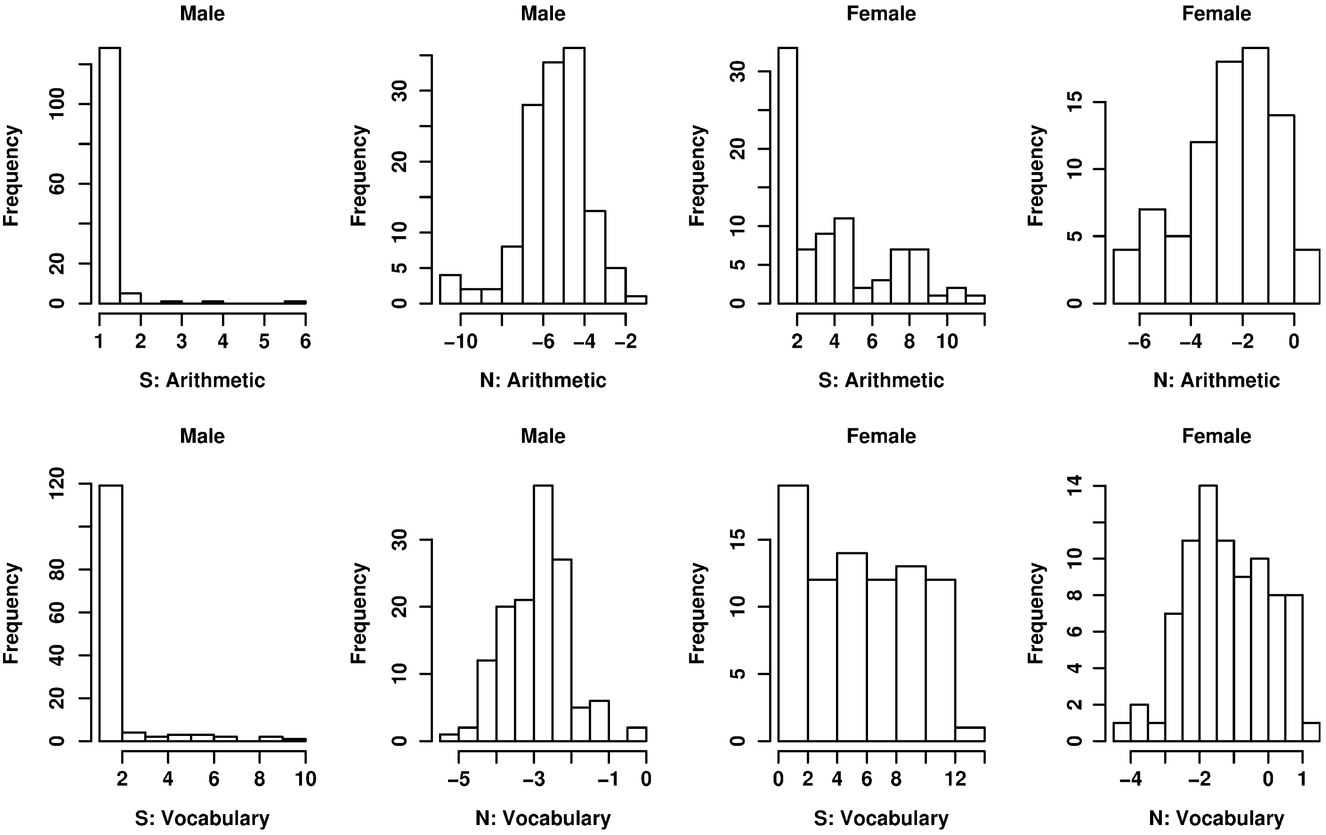

Fig. 1 Histograms showing distribution of standardized and normalized scores for representative subtests Arithmetic and Vocabulary, by sex (*S* standardized, *N* normalized)

scores and these are apparent from Fig. 1, as also described above, although to a lesser extent for females. Note that the interpretation of the normalized scores is that they are standard deviation units away from the general population mean (for a specific age band). As expected, for an ID population, as in the FXS population, the mean was negative and not symmetric about zero.

The average profile of the normalized subtest scores for the FXS study cohort is displayed in Fig. 2. On average, the FXS cohort (males and females combined) performed worst on Arithmetic and best on the Similarities and Information subtests. With respect to the Arithmetic subtest, the group was about 4.3 standard deviations below the general population mean and about 2.1 standard deviations below the general population mean for Similarities and Information subtests (Fig. 2). All other subtests ranged between 2 to 3 standard deviations below general population mean (Fig. 2). Similar profiles were observed with FXS males and females with normalized scores ranging between −2 to −1 standard deviations across subtests (except for Arithmetic) for females and scores between −4 to −3 standard deviations for males.

Figure 3 provides two specific case examples of 14-year-old boys with FXS, "John" and "Max" to illustrate differences between the usual standardized scores and the normalized scores derived in the study. John is a boy with a fully-methylated full mutation and 2.5% FMRP, VABS ABC standard score of 20, and FSIQ of 40. Max has repeat size mosaicism, 17% FMRP, and a VABS ABC of 38 and FSIQ of 40. As can be seen in the figure, both boys obtained standardized scores of 1 on all subtests with no variability (overlapping horizontal lines at top of figure). In contrast, the normalized scores demonstrated increased variability within each case, and somewhat lower scores for John, who had lower FMRP and adaptive behavior.

Correlation of normalized and standardized subtest scores to Vineland and FMRP

The correlation/association between the clinical outcome, Vineland ABC, and normalized scores for each subtest was stronger with higher (positive) point estimates than with standard scores (Table 3, Combined data). The correlations ranged from a low of 0.58 to a high of 0.80 for Object Assembly and Information normalized scores, respectively (combined data; all correlations, $p<0.001$). For males, correlation analysis without the floored values lead to appreciable loss of data, reduced power and many correlation estimates not statistically different from zero (e.g. Block Design, Comprehension, Object assembly, Picture Arrangement, Picture Completion, Similarities and Vocab-

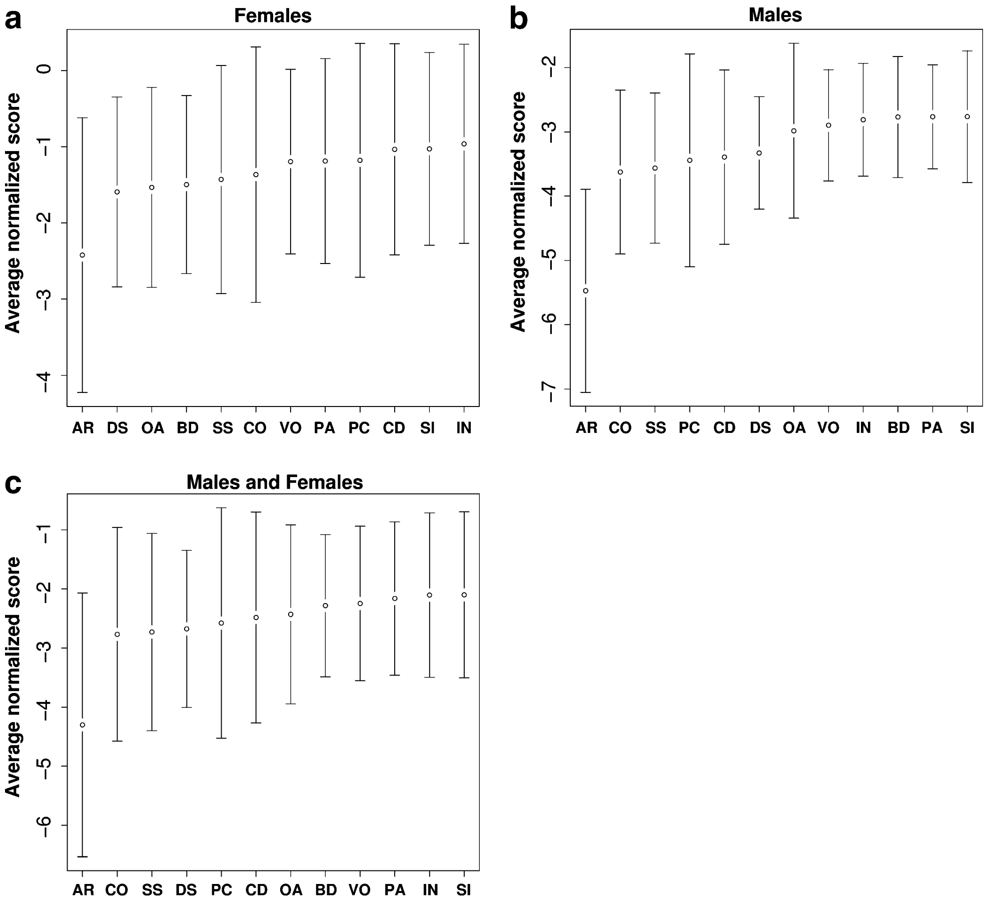

Fig. 2 Profile of normalized subtest scores in children with FXS (mean±SD). Subtests are ranked from lowest to highest. *IN* Information, *SI* similarities, *AR* arithmetic, *VO* vocabulary, *CO* comprehension, *DS* digit span, *PC* picture completion, *CD* coding, *PA* picture arrangement, *BD* block design, *OA* object assembly, *SS* symbol search

ulary; Table 3, male data). Interestingly, in females with FXS, where the effect of flooring was less due to reduced disease severity, point estimates for correlations between normalized scores with Vineland were higher than corresponding subtest estimates based on standardized scores (Table 3, female data). See Fig. 4 scatterplots displaying representative associations between standard versus normalized scores and Vineland ABC (Scatterplots for all subtests can be found in the supplemental materials).

A similar pattern of positive association was found with normalized subtest scores and FMRP, although the overall strength of association was weaker relative to Vineland ABC scores (Table 4, Combined data, all correlations, $p<$ 0.001). No association between standardized scores (without flooring) and FMRP was observed for all subtest scores for males, except for Digit Span, Picture Arrangement and Picture Completion. Similar to the pattern of association with Vineland score described above, the correlation estimates with normalized scores (significantly different from zero correlation) were observed more broadly across subtests in males (Table 4, male data). In females, no association was observed between FMRP and standardized score across all subtests, although stronger and significant associations based on normalized scores were observed

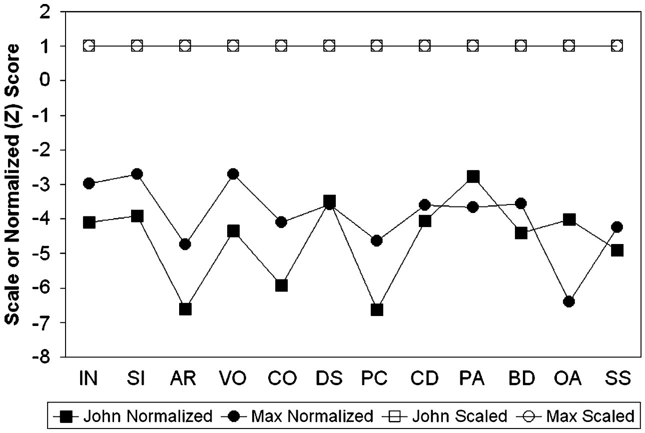

Fig. 3 Subtest standardized scores derived from the norms tables (*open symbols*) and normalized (*z*) scores (*closed symbols*) for two 14-year-old males with FXS. “John” has the full mutation and complete methylation, with 2.5% FMRP, Vineland Composite of 20, and FSIQ of 40. “Max” has repeat size mosaicism, 17% FMRP, Vineland Composite of 38, and FSIQ of 40. For the WISC-III, standardized scores have a range of 1 to 19, with a mean of 10 and standard deviation of 3. *IN* Information, *SI* similarities, *AR* arithmetic, *VO* vocabulary, *CO* comprehension, *DS* digit span, *PC* picture completion, *CD* coding, *PA* picture arrangement, *BD* block design, *OA* object assembly, *SS* symbol search

Table 3 Correlation of standard and normalized scores with Vineland Adaptive Behavior Composite

Subtest domain	Males and females		Males		Females	
	Standard score[a,b]	Normalized score[b]	Standard score[a]	Normalized score	Standard score[a]	Normalized score
Arithmetic	0.49	0.69	–	0.33**	0.46*	0.59**
Block design	0.42	0.67	−0.09	0.52**	0.33*	0.49**
Coding	0.61	0.76	0.45*	0.58**	0.49**	0.59**
Comprehension	0.62	0.77	0.29	0.61**	0.54**	0.62**
Digit span	0.55	0.71	0.41*	0.50**	0.43*	0.53**
Information	0.67	0.80	0.35*	0.64**	0.60**	0.66**
Object assembly	0.43	0.58	0.01	0.47**	0.37*	0.41**
Picture arrangement	0.51	0.71	0.11	0.51**	0.43*	0.53**
Picture completion	0.46	0.69	0.11	0.52**	0.23	0.54**
Similarities	0.50	0.70	0.10	0.50**	0.42*	0.52**
Symbol search	0.49	0.73	0.51*	0.52**	0.29*	0.59**
Vocabulary	0.52	0.72	0.15	0.49**	0.42*	0.53**

[a] Without floored value of 1. For males, sample sizes for these correlations ranged from 29 (Block Design) to 50 (Information). For females, sample sizes ranged from 54 (Arithmetic) to 73 (Information).

[b] All correlation values (standard and normalized scores) for combined data significantly different from zero ($P<0.001$)

** $P<0.001$

* $P<0.01$ or $P<0.05$

– Insufficient data

across several subtests (Table 4, female data). See Fig. 5 scatterplots displaying representative associations between standard versus normalized scores and FMRP (Scatterplots of all subtests can be found in the supplemental materials).

Mean normalized scores and IQ

Figure 6 displays the relationship between mean normalized scores versus standardized IQ scores with VABS ABC.

Fig. 4 Bivariate relationship/association of standardized vs. normalized score for representative subtests Arithmetic and Vocabulary with Vineland Composite score by sex (*red* female)

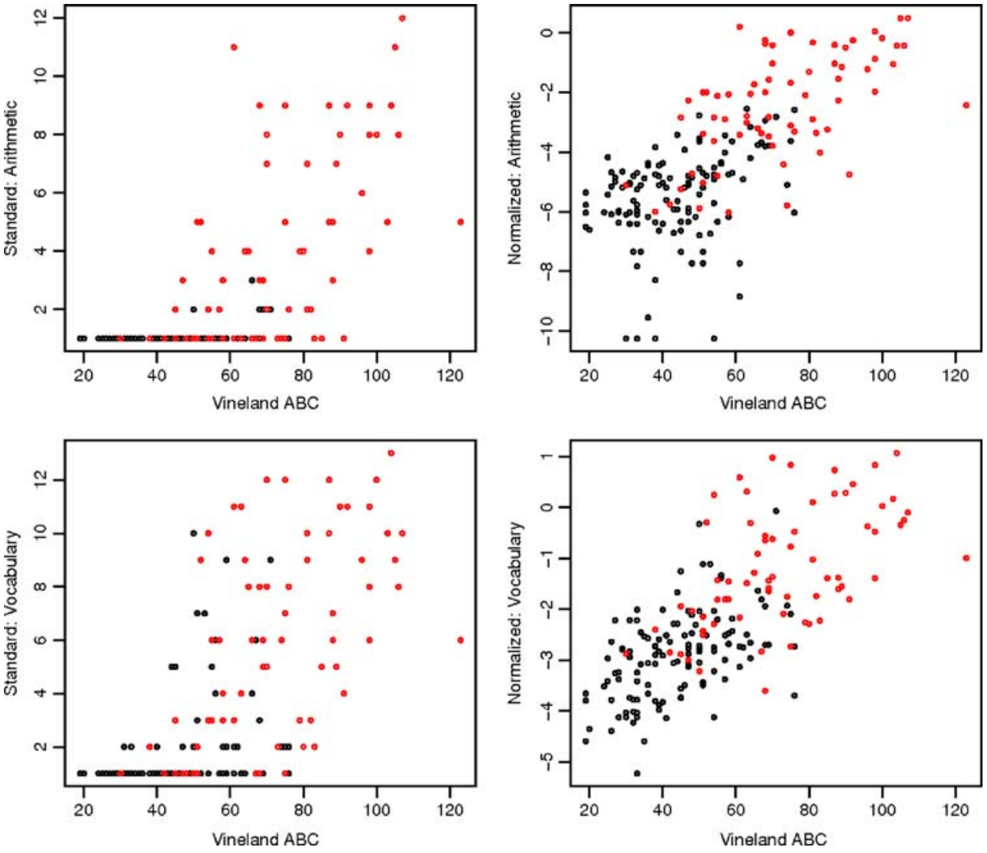

Table 4 Correlation of standard and normalized scores with FMRP

Subtest domain	Males and females		Males		Females	
	Standard score[a,b]	Normalized score[b]	Standard score[a]	Normalized score	Standard score[a]	Normalized score
Arithmetic	0.32	0.63	–	0.12	0.17	0.36*
Block design	0.35	0.47	0.04	0.09	0.16	0.20
Coding	0.35	0.60	0.01	0.24*	0.18	0.16
Comprehension	0.39	0.58	−0.07	0.17	0.26	0.28*
Digit span	0.44	0.64	0.64**	0.30*	0.15	0.30*
Information	0.45	0.63	0.18	0.21*	0.15	0.27*
Object assembly	0.35	0.41	0.00	0.06	0.26	0.19
Picture arrangement	0.44	0.57	0.35*	0.23*	0.20	0.23
Picture completion	0.53	0.58	0.36*	0.24*	0.15	0.23
Similarities	0.35	0.54	0.02	0.19*	0.17	0.20
Symbol search	0.42	0.61	0.36	0.27*	0.26	0.27*
Vocabulary	0.37	0.60	−0.13	0.21*	0.19	0.26*

[a] Without floored value of 1. For males, sample sizes for these correlations ranged from 25 (Block Design) to 46 (Information and Object Assembly). For females, sample sizes ranged from 59 (Arithmetic) to 76 (Information).

[b] All correlation values (standard and normalized scores) for combined data significantly different from zero ($P<0.001$)

** $P<0.001$

* $P<0.01$ or $P<0.05$

– Insufficient data

Similar to individual subtest standardized scores, IQ was predominantly floored, especially for males. Because subtest (standardized) scores for females were typically not floored across subtests, the IQ computed from these standard subtest standard scores, its association/correlation with Vineland outcome tracked closely that of corresponding correlation of Vineland to the normalized scores. This similarly held with correlation to FMRP (see supplementary materials).

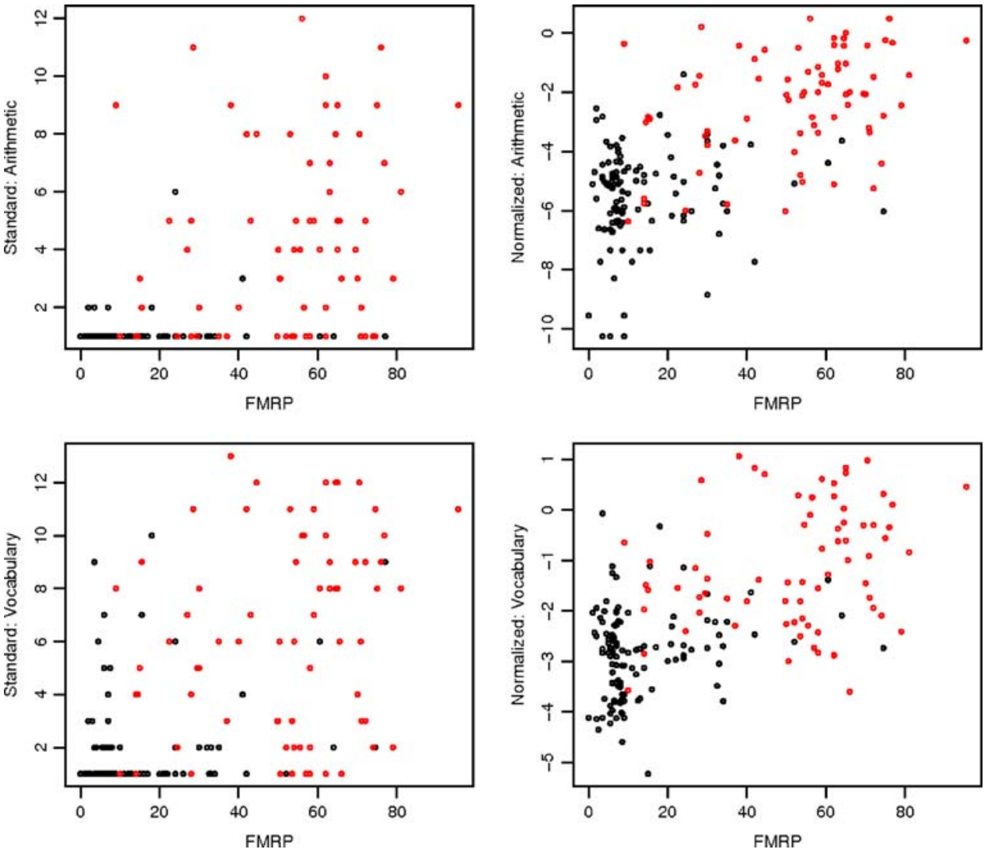

Fig. 5 Bivariate relationship/association of normalized vs. standardized score for representative subtests Arithmetic and Vocabulary with FMRP levels by sex (*red* female)

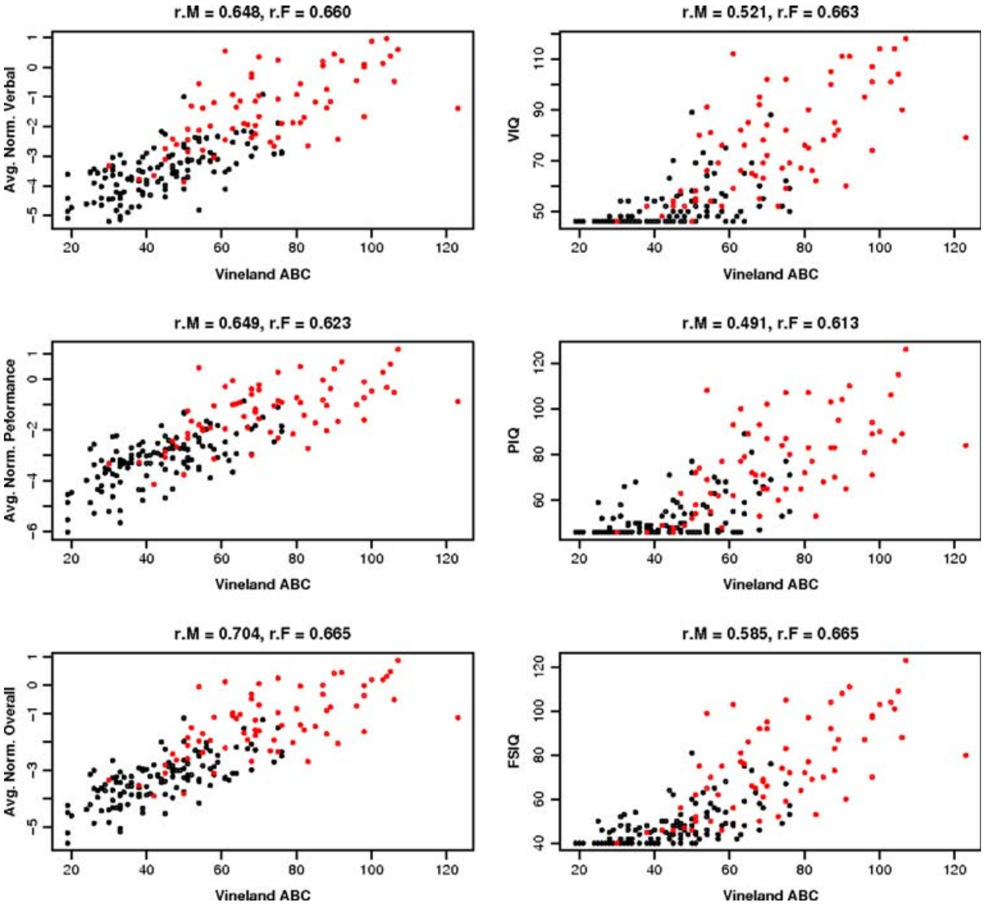

Fig. 6 Bivariate relationship/association of Vineland Composite with average normalized score (*left column*) vs. IQ (*right column*) by sex (*red* female). *M* Male, *F* female, *VIQ* verbal IQ, *PIQ* performance IQ, *FSIQ* full scale IQ

Comparison of intellectual ability classifications by FSIQ vs. mean normalized score

Participant WISC-III scores were classified by intellectual ability according to the DSM-IV guidelines for FSIQ and by analogous mean normalization scores, separately for males and females. As can be seen in Tables 5 and 6, approximately 85% of males with FXS had FSIQ scores within the Moderate range (owing primarily to the flooring IQ effect at 40). The mean normalized scores, however, show a much broader distribution with scores of several boys falling within the Profound range, and a large proportion changing classification from the Moderate to Severe range and from the Moderate to Mild range. The classification differences for females were less substantial; however it is notable that 20.5% of females had mean

Table 5 Intellectual ability classifications[1] (% of participants) determined by full scale IQ (FSIQ) versus mean normalized score

Classification[a]	Males		Females	
	Based on FSIQ (%)	Based on mean normalized (%)	Based on FSIQ (%)	Based on mean normalized (%)
Profound	0.0	3.3	0.0	0.0
Severe	0.0	17.0	0.0	2.2
Moderate	84.8	46.4	20.8	8.7
Mild	10.8	29.4	29.2	16.3
Borderline	3.8	1.3	15.6	26.1
Low average	0.6	2.6	10.4	21.7
Average	0.0	0.0	20.8	22.8
High average	0.0	0.0	3.1	2.2

[a] Classifications: profound, FSIQ<25 or mean $z<-5.0$; Severe, FSIQ≥25 and < 40 or mean $z\geq-5.0$ and < -4.0; Moderate, FSIQ≥40 and < 55 or mean $z\geq-4.0$ and < -3.0; Mild, FSIQ≥55 and < 70 or mean $z\geq-3.0$ and < -2.0; Borderline, FSIQ≥70 and < 80 or mean $z\geq-2.0$ and < -1.3; Low Average, FSIQ≥80 and < 90 or mean $z\geq-1.3$ and < -0.7; Average, FSIQ≥90 and < 110 or mean $z\geq-0.7$ and < 0.7; High Average, FSIQ≥110 and < 120 or mean $z\geq0.7$ and < 1.3. z score refers to the number of standard deviations from the average of the WISC-III normative sample.

Table 6 Differences in intellectual ability classification based on WISC-III full scale IQ (FSIQ) and classification based on mean normalized scores in boys and girls with FXS

FSIQ classification	Normalized classification	Males (%)	Females (%)
No change in classification		57.9	54.5
Moderate ID	Profound ID	3.3	0.0
Moderate ID	Severe ID	16.4	4.2
Moderate ID	Mild ID	19.1	10.2
Moderate ID	Borderline	0.7	20.5
Borderline	Low average	2.0	9.1
Low average	Average	0.0	1.1
Average	High average	0.0	1.1
High average	Average	0.0	1.1

normalized scores within the Borderline range whereas their FSIQ scores were in the mild range.

Discussion

The results of this study, using the Wechsler Intelligence Scale for Children, highlight significant floor effects and restricted sensitivity as major limitations of standardized intelligence testing of children with fragile X syndrome, one of the most common causes of intellectual disability. Despite the preponderance of floored standardized scores (up to 70% of the sample), we demonstrate that substantial and meaningful variability in performance of lower functioning individuals is lost in the standardization of raw scores. We show that renormalized scores that are based on the individual's actual deviation from the test normative data have a distribution and variability that is very much improved over the typical subtest standardized scores derived from norms tables. We show that relative to the usual subtest standardized scores, these normalized scores demonstrate more robust linear associations with a clinical measure of adaptive behavior (Vineland Adaptive Behavior Scale) and a genetic measure specific to FXS indicating the degree of *FMR1* protein deficiency. The normalized scores appear to provide a profile of relative strengths and weaknesses in lower functioning individuals that is not reflected in the usual standardized scores. On a group level, the normalized scores show a substantial deficit on the Arithmetic subtest, which is consistent with prior research highlighting this aspect of the FXS cognitive phenotype and its neuroanatomical basis [26]. These results appear to have major clinical and research implications for intelligence testing of children with FXS and probably other types of ID. Although we have only documented this problem in one population and with one intelligence test, the WISC-III, the results suggest that cognitive tasks that are integral to the measurement of IQ can be sensitive to individual differences, even in very low functioning individuals.

The results of this study also have important research implications. IQ is an almost universal variable in developmental, neuroscience and genetic studies as an outcome of interest, a predictor variable, or as a critical tool for group matching. The use of IQ in lower functioning individuals, as currently derived by standardized tests, in such studies appears to lead to poor estimates of true level of cognitive ability and potential, an "even" profile that may obscure significant relative strengths and weaknesses, lower estimates of associations with other behavioral, biological, and genetic measures of interest, and samples that are inadequately matched on this dimension. Indeed, in FXS and perhaps other neurodevelopmental disorders, it will become increasingly important to utilize sensitive cognitive tests for tracking change as new targeted treatment trials are implemented.

The renormalization and improved sensitivity of intelligence testing for individuals with ID has implications for future research on the neuropsychology, neuroimaging, and genetic bases of neurodevelopmental disorders. For example, cognitive phenotyping studies and other research programs aimed at establishing links between genotype and specific cognitive patterns would greatly benefit from using individual scores that more accurately reflect the true deviation from normal as well as relative strengths and weaknesses. This has immediate implications for fragile X research as we develop and validate much more accurate measures of FMRP expression that could ultimately be used as prognostic indicators of developmental trajectory. In neuroimaging studies, efforts to determine the impact of brain morphological and functional abnormalities on neuropsychological deficits or relatively preserved abilities would also depend on cognitive scores that reflect the ability of individuals with ID (often in the experimental group) as accurately as those with typical development (often in the control group). Finally, from a study design perspective, it is important for many clinical studies of individuals with neurodevelopmental disorders to include comparison groups that are well-matched on cognitive ability so that results can be more confidently attributed specifically to the disorder in question and not confounded by more general developmental differences. A more accurate estimate of cognitive ability, as is presented here, would lead to improved matching and more powerful research designs. We emphasize that the concepts and methodological/statistical approaches proposed here may impact our ability to find other links between behavioral or cognitive phenotypes and biomarkers/genotypes.

Although children with ID represent a small proportion of the population, they should receive intellectual assessments that are as sensitive and valid as those available to children who are higher functioning. Many intelligence tests currently report performance of children in special categories, such as those with mental retardation, autism, or specific learning disabilities; however these data are primarily for validation study and are separate from the normative sample. An ambitious but worthwhile solution to the sensitivity problem is to over-sample children who are lower functioning in the standardization studies and include tasks that can be completed across a broader range of developmental levels, including items designed for children with a mental age extending into toddlerhood. An over-sampling of these children would yield enough normative data from children of varying levels of impairment, allowing a lower IQ floor. In the meantime, the publishers of widely-used standardized tests should consider releasing the raw data obtained from their standardization samples into the public domain so that more accurate estimates might be derived for lower functioning individuals, at least in research applications.

In summary, we show significant floor effects and lack of sensitivity of IQ measurement in children with FXS and mental impairment that can be substantially ameliorated by calculating each child's actual deviation from the normative sample. The validity of this approach was accomplished by our demonstration of stronger associations between these new normalized scores and another measure of development and a genetic measure specific to FXS, in contrast to similar correlations with the traditional standardized scores. We hope that our observations and conclusions will lead to future studies examining the sensitivity of intelligence testing in other populations of children with neurodevelopmental disorders and to improved tools for measuring cognitive abilities and patterns of strengths and weaknesses in lower functioning individuals.

Acknowledgements This research was supported by National Institutes of Health Grants HD36071, HD02274, MH77554, MH64708, and MH50047 and Interdisciplinary Research Consortium grants (RL1 AG032115; RL1 AG032119; UL1 DE19583) from the National Institute of Dental and Craniofacial Research, and grant UL1 RR024992 from the National Center for Research Resources (NCRR), a component of the NIH Roadmap for Medical Research. The authors thank The Psychological Corporation for providing the essential WISC-III standardization raw data for this work. The first author (D.H.) is especially grateful to Bronwyn Glaser for fruitful discussions about cognitive testing in children with fragile X syndrome that set the stage for this project.

References

1. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. Washington, DC: American Psychiatric Association; 1994.
2. Sparrow S, Balla D, Cicchetti D. Vineland adaptive behavior scales—interview edition. Circle Pines, MN: American Guidance Service, Inc.; 1984.
3. Roid G. Stanford Binet intelligence scales. 5th ed. Rolling Meadows, IL: Riverside Publishing; 2003.
4. Elliott C. Differential ability scales. San Antonio, TX: Pearson; 2008.
5. Facon B. A cross-sectional test of the similar-trajectory hypothesis among adults with mental retardation. Res Dev Disabil 2008;29 (1):29–44.
6. Couzens D, Cuskelly M, Jobling A. The Stanford Binet Fourth Edition and its Use with individuals with Down syndrome: Cautions for clinicians. Int J Disabil Dev Educ 2004;51(1):39–56.
7. Devys D, Lutz Y, Rouyer N, Bellocq JP, Mandel JL. The FMR-1 protein is cytoplasmic, most abundant in neurons and appears normal in carriers of a fragile X premutation. Nat Genet 1993;4 (4):335–40.
8. Tamanini F, Willemsen R, van Unen L, Bontekoe C, Galjaard H, Oostra BA, Hoogeveen AT. Differential expression of FMR1, FXR1 and FXR2 proteins in human brain and testis. Hum Mol Genet 1997;6(8):1315–22.
9. Comery TA, Harris JB, Willems PJ, Oostra BA, Irwin SA, Weiler IJ, Greenough WT. Abnormal dendritic spines in fragile X knockout mice: maturation and pruning deficits. Proc Natl Acad Sci USA 1997;94(10):5401–4.
10. Galvez R, Greenough WT. Sequence of abnormal dendritic spine development in primary somatosensory cortex of a mouse model of the fragile X mental retardation syndrome. Am J Med Genet A 2005;135(2):155–60.
11. Irwin SA, Galvez R, Greenough WT. Dendritic spine structural anomalies in fragile-X mental retardation syndrome. Cereb Cortex 2000;10(10):1038–44.
12. Irwin SA, Idupulapati M, Gilbert ME, Harris JB, Chakravarti AB, Rogers EJ, Crisostomo RA, Larsen BP, Mehta A, Alcantara CJ, Patel B, Swain RA, Weiler IJ, Oostra BA, Greenough WT. Dendritic spine and dendritic field characteristics of layer V pyramidal neurons in the visual cortex of fragile-X knockout mice. Am J Med Genet 2002;111(2):140–6.
13. Irwin SA, Patel B, Idupulapati M, Harris JB, Crisostomo RA, Larsen BP, Kooy F, Willems PJ, Cras P, Kozlowski PB, Swain RA, Weiler IJ, Greenough WT. Abnormal dendritic spine characteristics in the temporal and visual cortices of patients with fragile-X syndrome: a quantitative examination. Am J Med Genet 2001;98(2):161–7.
14. Dyer-Friedman J, Glaser B, Hessl D, Johnston C, Huffman LC, Taylor A, Wisbeck J, Reiss AL. Genetic and environmental influences on the cognitive outcomes of children with fragile X syndrome. J Am Acad Child Adolesc Psychiatry 2002;41:237–44.
15. Glaser B, Hessl D, Dyer-Friedman J, Johnston C, Wisbeck J, Taylor A, Reiss A. Biological and environmental contributions to adaptive behavior in fragile X syndrome. Am J Med Genet 2003;117A(1):21–9.

16. Hessl D, Dyer-Friedman J, Glaser B, Wisbeck J, Barajas RG, Taylor A, Reiss AL. The influence of environmental and genetic factors on behavior problems and autistic symptoms in boys and girls with fragile X syndrome. Pediatrics 2001;108:E88.
17. Hall SS, Burns DD, Lightbody AA, Reiss AL. Longitudinal changes in intellectual development in children with fragile x syndrome. J Abnorm Child Psychol 2008;36(6):927–39.
18. Loesch DZ, Huggins RM, Hagerman RJ. Phenotypic variation and FMRP levels in fragile X. Ment Retard Dev Disabil Res Rev 2004;10(1):31–41.
19. Reiss AL, Freund LS, Baumgardner TL, Abrams MT, Denckla MB. Contribution of the FMR1 gene mutation to human intellectual dysfunction. Nat Genet 1995;11(3):331–4.
20. Wechsler D. Wechsler Intelligence Scale for Children, Third Edition. San Antonio, TX: Psychological Corporation; 1991.
21. Taylor AK, Safanda JF, Fall MZ, Quince C, Lang KA, Hull CE, Carpenter I, Staley LW, Hagerman RJ. Molecular predictors of cognitive involvement in female carriers of fragile X syndrome [see comments]. Jama 1994;271(7):507–14.
22. Tassone F, Hagerman RJ, Ikle DN, Dyer PN, Lampe M, Willemsen R, Oostra BA, Taylor AK. FMRP expression as a potential prognostic indicator in fragile X syndrome. Am J Med Genet 1999;84(3):250–61.
23. Willemsen R, Mohkamsing S, de Vries B, Devys D, van den Ouweland A, Mandel JL, Galjaard H, Oostra B. Rapid antibody test for fragile X syndrome. Lancet 1995;345(8958):1147–8.
24. Willemsen R, Smits A, Mohkamsing S, van Beerendonk H, de Haan A, de Vries B, van den Ouweland A, Sistermans E, Galjaard H, Oostra BA. Rapid antibody test for diagnosing fragile X syndrome: a validation of the technique. Hum Genet 1997;99 (3):308–11.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 1995;57:289–300.
26. Rivera SM, Menon V, White CD, Glaser B, Reiss AL. Functional brain activation during arithmetic processing in females with fragile X Syndrome is related to FMR1 protein expression. Hum Brain Mapp 2002;16(4):206–18.