Development and Psychopathology (2021), **33**, 727–738
doi:10.1017/S0954579420001480

CAMBRIDGE UNIVERSITY PRESS

Special Issue Article

Choices, challenges, and constraints: a pragmatic examination of the limits of mental age matching in empirical research

N. Russo[1], E. A. Kaplan-Kahn[1], J. Wilson[1], A. Criss[1] and J. A. Burack[2]

[1]Department of Psychology, Syracuse University, Syracuse, NY, USA and [2]Department of Educational and Counselling Psychology, McGill University, Montreal, Quebec, Canada

Abstract

The work of Ed Zigler spans decades of research all singularly dedicated to using science to improve the lives of children facing different challenges. The focus of this article is on one of Zigler's numerous lines of work: advocating for the practice of mental age (MA) matching in empirical research, wherein groups of individuals are matched on the basis of developmental level, rather than chronological age. While MA matching practices represented a paradigm shift that provided the seeds from which the developmental approach to developmental disability sprouted, it is not without its own limits. Here, we examine and test the underlying assumption of linearity inherent in MA matching using three commonly used IQ measures. Results provide practical constraints of using MA matching, a solution which we hope refines future clinical and empirical practices, furthering Zigler's legacy of continued commitment to compassionate, meaningful, and rigorous science in the service of children.

Keywords: developmental approach, intellectual disability, IQ, mental age matching

(Received 28 July 2020; accepted 30 July 2020)

"The I.Q. is only a rate measure in the sense that it relates a non-psychological measure (passage of time) to a psychological one (level of cognition achieved). Approached in this way it is the MA (level) and not the I.Q. (the relationship of MA to chronological age) that determines the exact nature, including the rate, of learning any task. (Zigler, 1967, p. 578)."

Originally, this paper, based on Ed Zigler's articulation of the essential need for mental age (MA) matching in empirical comparisons between persons with intellectual disability and typically developing children, was meant to include four generations of scholars beginning with Ed, who mentored Jake Burack, who mentored Natalie Russo, who mentored Elizabeth Kaplan-Kahn. It also includes colleagues from computational cognitive science who, at a seminar, asked the question that is at the core of this paper, and then helped us to address this essential methodological issue in the developmental approach to intellectual disability. Although Ed passed away before we began writing this paper so cannot be included as an author, we hope that it is consistent with his basic value that science is only meaningful when it is used to help others and that it contributes to both research and clinical work focused on individuals with intellectual disability. The questions we aim to answer, which we will frame in terms of the evolution of arguments of the developmental approach to developmental disability are: when do our methods for MA matching work? and when do they not? The goal of this paper is to answer this question in the most pragmatic sense of the term. That is, at which point does the relationship between MA, chronological age (CA), and IQ change across different IQ tests? The answer to this question provides computational limits on how we use IQ tests to compare the performance of those with and without intellectual disability. The intergenerational tracing of this question starts with the publication of Ed's (Zigler, 1967, 1969) early papers on the topic.

The Second Normal Distribution

With the impressive advances in biologically based research generally, and in genetic research specifically (Chiurazzi & Pirozzi, 2016; Mir & Kuchay, 2019; Wolfe, Strydom, & Bass, 2019), the number of identified genetic causes of disorders associated with intellectual disability has grown exponentially (Abbeduto, Thurman, Bullard, Nelson, & McDuffie, 2019; Vissers, Gilissen, & Veltman, 2016; Vorstman & Ophoff, 2013), and so too have the attempts to understand their nosology (e.g., Stevenson, 2000), etiology (e.g., Iwase et al., 2017; Karam et al., 2016), and prognosis (e.g., Hanaoka et al., 2010; Katz & Lazcano-Ponce, 2008). While biological advances have provided clues related to the origins of more than 1,000 disorders that impact cognitive function, the field of psychology and its allied disciplines has been focused more on the "now what" questions. Now that we know the cause, how do we understand the effect? For example, we know that atypical cell division which results in a third chromosome 21 leads to a condition called Down syndrome. The role that psychology has played is in

Author for Correspondence: Natalie Russo, Associate Professor, Department of Psychology, Syracuse University, 430 Huntington Hall, Syracuse, NY, 13244; E-mail: nrusso@syr.edu

Cite this article: Russo N, Kaplan-Kahn EA, Wilson J, Criss A, Burack JA (2021). Choices, challenges, and constraints: a pragmatic examination of the limits of mental age matching in empirical research. *Development and Psychopathology* **33**, 727–738. https://doi.org/10.1017/S0954579420001480

understanding the specific impacts of this third chromosome on the cognitive, behavioral, social, and emotional life of those with the condition as well as on their family members.

At the core of this work is the developmental approach to intellectual disability that was originally delineated by Ed and colleagues (Hodapp, Burack, & Zigler, 1990; Zigler, 1967; Zigler & Balla, 1982; Zigler & Hodapp, 1986) and extended to include persons with organic etiologies by Dante Cicchetti and colleagues (Cicchetti & Beeghly, 1990; Cicchetti & Ganiban, 1990; Cicchetti & Pogge-Hesse, 1982). This approach focuses on how we study those with intellectual disability in order to gain a meaningful understanding of the development of their cognitive, behavioral, social, and affective functioning. Who do we make comparisons with, and on what basis are we comparing? Students of the developmental approach are familiar with these questions, and much has been written about their essential role in research with populations of individuals with intellectual disabilities (Burack, Dawkins, Stewart, Iarocci, & Russo, 2012a; Burack, Iarocci, Flanagan, & Bowler, 2004; Burack, Russo, Flores, Iarocci, & Zigler, 2012b; Flanagan, Russo, Flores, & Burack, 2008; Jarrold & Brock, 2004; Mervis & Klein-Tasman, 2004; Mervis & Robinson, 1999). However, the failure to consider these essential developmental issues continue to plague and fatally flaw many well-meaning and even well-funded empirical efforts (Burack, Russo, Gordon Green, Landry, & Iarocci, 2016b). Initially, Ed and colleagues referred to those who failed to adequately consider MA as defect theorists (Bennett-Gates & Zigler, 1998; Hodapp & Zigler, 1995; Zigler, 1967, 1969; Zigler & Balla, 1982), but concerns regarding the confounded and misleading findings have recently been reignited with the advent of researchers using neuroscience techniques (Burack et al., 2016b) to examine brain–behavior relationships in populations of individuals with intellectual disabilities.

In a classic example, defect scientists of the 1960s were focused on understanding the key underlying deficit of various intellectual disabilities, with each touting their area of research as fundamentally causal to the cognitive dysfunction of a group (Zigler, 1967, 1969). In so doing, scientists compared performance on tasks of their construct of interest (e.g., attention, executive function, memory) between individuals with intellectual disability and those with average IQs of the same CA (for a discussion, see Burack et al., 2016b). When they inevitably found that individuals with intellectual disability performed worse on the specific task at hand in comparison to a group of individuals who had better overall cognitive function, they touted their area or construct as central or causal of the target group's intellectual difficulties. Of course, however, what they were finding was that a group of individuals with lower cognitive abilities performed worse in a particular area of function that was highly reliant on cognition in relation to those with higher cognitive abilities (Burack, Cohene, & Flores, 2011; Burack, Evans, Klaiman, & Iarocci, 2001; Burack et al., 2012b; Iarocci & Burack, 1998). Clearly, making comparisons between groups of individuals at the same CA, but, by definition of one group's disability status, at different levels of cognitive development precluded any scientifically sound conclusions about a specific area of cognition being meaningfully delayed beyond the a priori general differences in level between the groups.

This research led to many problematic and scientifically flawed claims that were translated into ineffective approaches to interventions and years of lost time with respect to science, practice, and education for the individuals and their families. While these fundamental flaws seem perhaps like an error "of the time," this method of making comparisons on CA, rather than on developmental level, still plagues our clinical and scientific landscape (Burack et al., 2016b). The implications even extend beyond work with persons with intellectual disability to any group with some consistently lower-than-average IQ levels. For example, in a study with common clinical tests, Lane et al. (2014) argued that attention processing appears to be commensurate with MA levels among children with fetal alcohol syndrome and related conditions for whom IQ is often lower than average, and that the common perceptions of attention deficits in this group are likely the consequence of the parent, teacher, and psychologist's expectations for the individuals' CA. Thus, in contrast to the defect theorists, the proponents of the developmental approach recast claims of deficit and defect into opportunities to understand developmental organization and coherence (Burack et al., 2016a; Cicchetti & Ganiban, 1990; Cicchetti & Pogge-Hesse, 1982; Hodapp et al., 1990).

The Developmental Approach to Intellectual Disability

The developmental approach has as one of its central tenets that one must consider development in order to understand intellectual disabilities. While this sounds like a rather simple truism, it is often overlooked and difficult to quantify from a research perspective. However, to truly understand the impact of intellectual disability on a particular psychological construct, one must consider the impact of the individual's experiences, their rate of cognitive growth, their history of successes and failures, their social interactions and their environment, and how this differs from typically developing individuals, or groups of individuals in the case of research. While theoretically this makes sense, it is practically and pragmatically an impossible task as developmental rates differ as a function of the skills being measured.

Accordingly, in operationalizing the developmental approach, a construct referred to as MA has been invoked. MA reflects the relationship between an individual's CA and level of skill (broadly defined). MA has been used to provide a reasonable proxy with which to compare the performance of a group of individuals with an intellectual disability to a group of typically developing individuals because using this construct attempts to account for the fundamental differences in these two groups' cognitive abilities. In this way, we can understand whether a particular area or skill level is commensurate with a child's developmental level while considering their overall cognitive capacity, or whether there are specific deficits or delays at hand that are uniquely related to a particular individual or group's phenotype.

Despite all its benefits, MA matching is not without some complications. Given their differing rate of development, the process of reaching any given MA is obviously longer for a person with intellectual disability. Thus, for example, using the traditional formula of (MA/CA) x 100 = IQ, a child with an intellectual disability and an IQ of 60 will be 10 years old when they attain the MA of 6, whereas their typically developing peer will be 6 years old. Clearly, in addition to the different experiences associated with their differences in intellectual level and associated conditions, the child with intellectual disability has lived considerably longer with all the inherent consequences of that fact. In addition, the rate of development is itself a critical issue to consider with regard to its meaning for the strength of the attainment of developmental milestones and for the sequelae of the moment in time of matching, as the MAs of children of

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

differing IQs that are the same at one moment in time will immediately diverge in the next moment.

MA matching, thus, is clearly not a magic bullet and cannot account for all aspects of development, but it is a better approach within the inherently imperfect enterprise of empirical work that allows for the delineation of strengths and weaknesses in areas of functioning relative to an individual's level of skill development (e.g., Campbell et al., 2013; Russo et al., 2007), and has served to debunk many of the defect theorists claims (Burack et al., 2001, 2016; Burack et al., 2012a).

Fundamental Theoretical Assumptions of Mental Age Matching

Between-group comparisons are inextricably woven into the very conceptualization of intellectual disability (Burack, 1997; Cicchetti & Pogge-Hesse, 1982; Zigler, 1967, 1969; Zigler & Hodapp, 1986). No matter what the variable of interest, whether it be cognitive, emotional, behavioral, or physical, the concept of intellectual disability is defined by a comparison to a "typically developing" distribution. As these comparisons are the marker with which we delineate our diagnostic boundaries, answers to the questions of *who* we are comparing, *which* measures we are using to make comparisons, and *how* we match the groups have the ability to powerfully shape our understanding of persons with intellectual disability.

Generally speaking, IQ tests are normed to provide a score that represents the relative standing of an individual in relation to other individuals of the same CA. Theoretically, IQ scores represent the rate at which someone has achieved a particular set of skills, in relation to the rate of skill acquisition of similarly aged peers. Commonly, IQ tests have a mean standard score of 100 and a standard deviation (*SD*) of 15. As such, scores higher than 100 indicate that an individual is acquiring expected skills faster than same aged individuals, and scores below 100 indicate that the relationship between skill acquisition and age is slower than average. That is, if we consider the numerator the level of skills acquisition of a child and the denominator the child's CA, the greater the numerator in relation to the denominator, the faster the developmental growth of the child and the higher their IQ score will be. For example, a child who attains the types of skills common to a 6-year-old in 6 years, and therefore has an IQ of 100, is developing more slowly than a child who takes 5 years to acquire the same skills and has an IQ of 120, and is developing more quickly than a child who took 8 years and has an IQ of 75.

IQ is thought to be stable after early childhood (Hoekstra, Bartels, & Boomsma, 2007; Schneider, Niklas, & Schmiedeler, 2014) through adolescence and also predicts adult levels of functioning (Koenen et al., 2009; McCall, 1977). However, while IQ scores can be used directly when comparing two groups of similar cognitive abilities, this solution is not tenable in the case where there is a discrepancy in skill acquisition rates between the two groups being compared, as is the case when conducting research comparing individuals with intellectual disabilities to typically developing children. By definition, individuals with intellectual disabilities acquire skills at slower rates than typically developing individuals. Thus, matching on the basis of IQ alone leads to a similar dilemma of matching on the basis of CA – by definition of their designations, the groups will never be comparable on this dimension.

The primary goal in matching clinical and neurotypical groups is to ensure that the groups are at a comparable developmental level (Burack, Iarocci, Bowler, & Mottron, 2002; Burack et al., 2004; Mervis & Klein-Tasman, 2004). This practice makes it possible to ask the empirically precise question: does the clinical group's development *of skill/ability X* differ from that of typical development? By nature of their diagnosis, the clinical group's general development differs from what is considered typical; however, in order to answer more rigorous questions regarding the development of a specific cognitive construct, rather than general development, it is necessary to, at the very least, attempt to equate general developmental level.

Even matching groups on overall IQ is problematic as it assumes that this single number, which reflects a weighted sum of verbal and nonverbal cognitive abilities, is acquired in the same way between populations, but this is not usually the case across developmental disorders with an etiological cause. Take, for example, the situation where groups of individuals with Williams syndrome (WS) are being compared to individuals with Down syndrome on the basis of overall IQ to assess differences in some construct such as nonverbal working memory. As a result of stronger verbal than nonverbal skills that are common among persons with WS (Bellugi, Bihrle, Jernigan, Trauner, & Doherty, 1990; Mervis & Klein-Tasman, 2000), the overall IQ scores which reflects a combination of scores on verbal and nonverbal abilities, of those with WS, would overestimate their abilities in a nonverbal domain, especially in comparison to a group of individuals with Down syndrome who have stronger nonverbal than verbal skills. That is, despite having the same overall IQ score, this single number does not accurately reflect the group differences in patterns of abilities that would meaningfully be linked to performance and might lead to an erroneous conclusion that a deficit in a particular area of function is present when it is not. That is, matching on global level may not have been an adequate matching strategy and instead a decision should have been made to match on the basis of the subtest that most closely matched the construct being measured (Burack et al., 2004; Mervis & Klein-Tasman, 2004), a task to which the data we present below can contribute. While matching approaches clearly are not a panacea, and strategies for adequate comparisons continue to be refined, they play an essential role in basic comparisons between groups.

The goals of the developmental approach to intellectual disability include understanding the cognitive strengths and weaknesses of a group of individuals who share similar etiologies. Findings of "true" delays or deficits of an area of function relative to what would be expected given that groups development provides researchers and clinicians with an opportunity to develop interventions to remediate skill differences. Thus, precise methodologies, and approaches are needed to determine where and how to best intervene to support the learning and function of those with intellectual disability. To ensure that our matching measures are adequate, they must be informed by both our theoretical understanding of development and the constructs we are measuring in our studies, as articulated in the developmental approach to intellectual disability, as well as by our understanding of the empirical limits of the tests we use to make matching decisions.

Fundamental Statistical Assumptions of Mental Age Matching

Fundamentally, matching strategies rely on the assumption that there is a linear relationship between a person's score on an IQ

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

test and their age. The complexity of matching occurs when there is a meaningful discrepancy between groups in the relationships between age and IQ. In this case, MA matching strategies can and should be used. MA is used frequently in comparison studies of, for example, individuals with an intellectual disability and typically developing participants. In essence, the question asked by using the strategy is: at what age would person X's raw score be considered average. It is here that the key assumption of linearity is invoked.

As an example of why linearity is a crucial assumption in the MA matching formula, consider two separate scenarios. In the first, MA matching may be used to match a 10-year-old with an IQ of 70 whose MA is 7 years (MA = 10*70/100) with a 7-year-old with an IQ of 100 whose MA is also 7 years (MA = 7*100/100). Researchers who adopt the developmental approach to intellectual disability research would likely support the comparison of these individuals on a particular cognitive construct (e.g., working memory) because the children would be approximately matched for developmental level. The MA-matching equation, however, would not likely be used to compare a 50-year-old with an IQ of 50 whose MA is 25 years (MA = 50*50/100) with a 25-year-old with an IQ of 100 whose MA is also 25 years (MA = 25*100/100). Here, it seems unreasonable to compare the working memory abilities of the 50-year-old with an IQ of 50 to that of the typically developing 25-year-old because we do not expect working memory abilities to continue to develop linearly through adulthood, and do not expect cognitive abilities to continue increasing with the same slope much past the ages of 16–18 years (e.g., Biggs & Collis, 1982; Case, 1980; Selman, 1980). Though extreme, this example illustrates the importance of linearity in MA matching. Linear development is largely assumed when matching across intelligence tests and age ranges, yet where this assumption holds, for which tests, at what ages, has, to our knowledge, yet to be empirically determined.

Three intelligence tests that are commonly administered in both research and clinical settings are the Wechsler Abbreviated Scale of Intelligence – second edition (WASI-II; Wechsler, 2011), the Wechsler Intelligence Scale for Children – fifth edition (WISC-V; Wechsler, 2014), and the Stanford Binet Intelligence Scales – fifth edition (SB5; Roid, 2003). These measures are intended to be administered individually to participants and have multiple subscales that assess different domains of intelligence. Each of the tests have different factor structures, and each of the original measure authors had different goals in constructing these scales, as well as different philosophies related to both what "intelligence" is and how we measure it.

When Binet first published the revision of the Binet–Simon test in 1908, expanding the number of items and providing age-based norms for children, he focused on the power of the overall score of the test (commonly understood now to reflect "*g*" or a general intelligence factor), stating that "it matters very little what the tests are so long as they are numerous" (Binet, 1911/1916, p. 329; as cited in Boake, 2002). In contrast to Binet's monothetic approach, Weschler, who began his career as a psychometrician in the Army administering the Army Alpha and Army Beta tests (used to rule out those who were and were not deemed competent to be in the war), noted that the strength of the Army tests were that one could analyze "the subject's performance on the individual tests which comprise the examination, in order to discover" if the subject had "any special abilities or disabilities" (Wechsler, 1932, p. 254; as cited in Boake, 2002). Based on his experience that subtests were important clues to abilities, his statistical critique that MA was not a valid manner in which to measure adult intelligence owing to the statistical artifacts of applying MA ratio to adults, he promoted the use of standard scores as a function of the mean deviation and *SD* of particular age groups.

There have been years of debates surrounding best practices in matching strategies, and "battles" between (a) defect theorists who invoke that intellectual disability is pathognomonic with cognitive deficits, justifying comparing those with an intellectual disability to typically developing children matched on CA, and (b) developmental psychopathologists who want to understand whether a particular skill is impaired or intact on the basis of an individual's developmental level. Nonetheless, despite these debates, fundamental, statistical assumptions that underlie developmental matching practices have not been verified. That is, is there a linear relationship between age and performance on IQ tests such that the MA = IQ*CA/100 formula provides an accurate level of comparison? Do these assumptions hold at some ages and not others? On some subtests and not others? For some IQ tests and not others? At what age does this relationship begin to break down? It is these set of questions which we set out to test using three commonly used standardized IQ tests in studies of individuals with intellectual disability: the WISC-V; the WASI-II, often used in research studies; and the SB5. In doing so, we rely on the normative data tables presented in the manuals of each of these tests, rather than on data that we have explicitly collected. Further, in the analysis section, we remain agnostic to the structure of the tests. That is, we treat all subscale and standard scores for each test in the same manner. We begin by describing the psychometric properties of the three tests we are examining, followed by a description of our process, estimation, parameterization, and analysis procedures. Our goals are to provide evidence for the empirical limits of all of the subscales of each test in line with Ed's essential call for MA matching in the study of persons with intellectual disability.

Method

Measures

Wechsler Abbreviated Scale of Intelligence – second edition (WASI-II)

The WASI-II was published in 2011 and is used to assess individuals ranging from 6 to 90 years old. It includes four subtests: block design, vocabulary, matrix reasoning, and similarities. Normative data for the WASI-II are based on a sample of 2,300 examinees that were tested between January 2010 and June 2011. One hundred examinees were tested in each of the 23 age groups of the WASI-II. The range of ages in each age group varies and spans from 7 months (i.e., 6:0–6:7) to 20 years (i.e., 45–64). Examinees in the normative data sample were stratified on key demographic variables including sex, race/ethnicity, self or parent education level, and geographic region based on 2008 United States census data.

The WASI-II Manual (Wechsler, 2011) reports adequate reliability and validity. Internal consistency reliability coefficients for each of the WASI-II subtests were obtained for both children and adult samples using the split-half method and range from .83 to .95. Test–retest stability was calculated using Pearson's product-moment correlation for each subtest by retesting a subsample of 215 participants 12–88 days after the original testing. Corrected stability coefficients range from .79 to .96, which are

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

considered high for test-retest reliability. Interscorer Agreement for the WASI-II was also high across all of the subtests, ranging from .84 to .99. Evidence for the validity of the WASI-II is based on the test content, internal structure, and relationship with other measures assessing the same or similar constructs. These measures all supported adequate levels of validity in the WASI-II, including significant correlations with other measures, such as the WISC-IV, the Wechsler Adult Intelligence Scale – fourth edition, and the Kaufman Brief Intelligence Test, which measure similar constructs to those measured by their WASI-II counterpart subscales.

Wechsler Intelligence Scale for Children – fifth edition (WISC-V)

The WISC-V was published in 2014 and is used to assess the intelligence of children ages 6 to 16 years old. It includes a total of 16 subtests: block design, similarities, matrix reasoning, digit span, coding, vocabulary, figure weights, visual puzzles, picture span, symbol search, information, picture concepts, letter-number sequencing, cancelation, comprehension, and arithmetic. Normative data for the WISC-V are based on a sample of 2,200 children that were tested between April 2013 and March 2014. Two hunderd examinees were tested in each of the 11 age groups of the WISC-V. Each of the 11 age groups of the WISC-V is 12 months (e.g., 6:0–6:11). Examinees in the normative data sample were stratified on key demographic variables including sex, race/ethnicity, parent education level, and geographic region based on 2012 United States census data.

The WISC-V Technical and Interpretative Manual (Wechsler, 2014) reports adequate reliability and validity. Internal consistency reliability coefficients for each of the WISC-V subtests were obtained for all age-grouping samples using the split-half method, corrected by the Spearman–Brown formula, and range from .81 to .94. Test–retest stability was calculated using Pearson's product–moment correlation for each subtest by retesting a subsample of 218 participants 9–82 days after the original testing. Corrected stability coefficients range from .71 to .90, which are considered adequate for test–retest reliability. Interscorer Agreement for the WISC-V was also high across all of the subtests, ranging from .97 to .99. Evidence for the validity of the WISC-V is based on the test content, internal structure, and relationship with other measures assessing the same or similar constructs. These measures all supported adequate levels of validity in the WISC-V, including significant correlations with other measures, such as the Wechsler Preschool and Primary Scale of Intelligence – Revised, the Wechsler Adult Intelligence Scale (fourth edition), and the Kaufman Assessment Battery for Children (second edition), which measure similar constructs to those measured by their WISC-IV counterpart subscales.

Stanford–Binet Intelligence Scales – fifth edition (SB5)

The SB5 was published in 2003 and is used to assess individuals who are 2 to 90 years old. It includes a total of ten subtests, with nonverbal and verbal subtests for each of the following five constructs: fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory. Normative data for the SB5 are based on a sample of 4,800 examinees that were tested over a 12-month period in 2001 and 2002. Thirty age groups were defined for the sampling purposes of the SB5, with the range of ages in each group varying between six months (i.e., 2:0–2:6) to 10 years (i.e., 40–49). Examinees in the normative data sample were stratified on key demographic variables including sex, race/ethnicity, socioeconomic level, and geographic region based on 2001 United States census data.

The SB5 Technical Manual (Roid, 2003) reports adequate reliability and validity. Internal consistency reliability coefficients for each of the SB5 subtests were obtained for all age-grouping samples using the split-half method, corrected by the Spearman–Brown formula, and range from .72 to .96. Test–retest stability was calculated using Pearson's product–moment correlation for each subtest by retesting a subsample of 356 participants 1–39 days after the original testing. Corrected stability coefficients range from .66 to .91, which are considered adequate for test–retest reliability. Interscorer Agreement for the SB5 was also high across all of the subtests, ranging from .74 to .97. Validity of the SB5 is demonstrated based on the assessment of content-related, criterion-related, and construct-related evidence. These measures all supported adequate levels of validity in the SB5, including significant correlations with other measures, such as the Wechsler Preschool and Primary Scale of Intelligence – Revised, the Wechsler Adult Intelligence Scale – third edition, the Woodcock-Johnson III Tests of Cognitive Abilities, and the Wechsler Individual Achievement Test – second edition, which measure similar constructs to those measured by their SB5 counterpart subscales.

Procedures

Test construction assumptions and known parameters

Different intelligence tests build IQ subscale norms for a population of a given age range with the same general procedure. In general, a subscale is created and standardized by testing a large number of people within an age bracket, then assigning a subscale score to each level of performance based upon the number of people who achieved that score or lower. The higher the subscale score, the fewer individuals who performed at least that well. For all three IQ tests considered in this article, the subscale scores on the tests are capped such that the lowest obtainable score is received only by those who are at most 3 *SD*s below the mean (the 0.1% percentile), and those who obtain the highest score perform at least 3 *SD*s above the mean (the 99.9% percentile). The WASI-II reports standardized scores as *t* scores with a mean of 50 and a *SD* of 10. The WISC-V and the SB5 report standardized scores as scaled scores with a mean of 10 and *SD* of 3. For the sake of simplicity, these two types of scores will be both referred to as *standard scores*. Standard scores can be summed and converted to create different composite scores for each assessment (e.g., Full-Scale Intelligence Quotient [FSIQ], Verbal Comprehension Index [VCI], or Perceptual Reasoning Index [PRI]). Because age is not a variable when converting a standard score sum to a composite score, the current article focuses on standard scores at the level of individual subscales. Table 1 gives the minimum, median, and maximum subscale scores for the three tests on the basis of the test manuals.

Estimations

To estimate the performance distribution for each age group on each subscale test, the process of assignment of scores to percentiles and standard scores (described above) was reversed. In a first step, standard scores were transformed into *z* scores based on the knowledge of the center of each scaled score and knowledge that the largest and the smallest standard score represents a performance (number of correct responses) 3 *SD*s above and below the mean, respectively. In a second step, *z* scores were converted into

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

Table 1. Minimum, median, and maximum subscale scores, by IQ test

	Subscale score		
IQ test	Minimum	Median	Maximum
SB5	1	10	19
WASI-II	20	50	80
WISC-V	1	10	19

Note. These values were obtained from the respective test manuals.

percentiles, providing a cumulative performance distribution. In a third step, this cumulative performance distribution was converted into a probability distribution, from which the mean deviation and *SD* were extracted. These derived means and *SD*s formed the basis of the main analysis. For example, if a fictional subscale on the Stanford–Binet stated that a raw score of 20 gave a standard score of 10 (the median), and a raw score of 21 gave a standard score of 11 (0.333 *SD*s above the mean), we could infer that 50% of people would score 20 or below and 63% of people would score 21 or below. Therefore, 13% (63%–50%) of the sample in the norming study received a raw score of 21. In this way, we recreated the distribution of raw scores for every age bracket. This process allowed us to consider the raw performance of each age bracket and test whether these scores increase linearly. Note that, because the rounding and truncating procedures used by the makers of IQ tests result in a loss of information, these means and *SD*s are estimates of the original distributions. However, the difference between those and the ones presented here are minimal.

The process described above was repeated for every subscale and every age range presented in each of the test manuals. Examples of these subscale score distributions are presented in Figure 2, which plots the mean deviation and *SD* of subscale scores by the center of each age bracket, excluding age brackets for those over 20 years old.

Analysis procedures

Having reverse-engineered the distributions of raw scores, we subsequently examined how performance on the IQ tests changes with age and assessed the veracity of the notion that MA = CA*IQ/100. Since adult IQ scores are stable, and as described in the example above, the linear relationship between age and IQ was not expected to hold in the oldest age brackets, we only considered the raw scores for participants less than 20 years of age. The aim of this analysis was to determine whether a simple linear model could be applied to the data. If different linear functions were necessary to describe the age–score relationship for different age ranges, then these different ranges could not be compared to each other using the MA-matching theory and, instead, some other method would be necessary.

We tested for linearity in each subscale by fitting multiple segmented regressions and choosing the best model using the Bayesian information criterion (BIC). Segmented regression is a technique to determine which regions of the data share the same relationship with some variable. The BIC combines the likelihood or degree of fit with model complexity to give a measure of quality. An ideal model would be extremely simple yet fit the data very well. Lower values of the BIC imply a higher quality model. We used the BIC to determine how many segments are optimal in the segmented regression. In other words, we used the BIC to determine how many disparate sets of linearly comparable ages there are for participants under age 20. If the best fit indicated a single segment, then the entire set of age brackets were best fit by a simple linear regression and all age brackets are linearly comparable to one another. If the best fit included multiple segments, then the linear function that best described the age–performance relationship changes with age, and some brackets could not be linearly compared to some others. For each subscale, we fit six regressions with 1, 2, 3, 4, 5, or 6 segments using the *segmented* package in R (Muggeo, 2008); the package uses a Taylor expansion of the regression to transform the nonlinear problem of where to place break points into a linear regression problem (Muggeo, 2003). To ensure an accurate fit and to ensure that the model can capture discontinuities in the data (two such discontinuities appear in the WISC-V for the Coding and Symbol Search subtests because they have different sets of norms for different age groups), we divided each age bracket into 10 smaller brackets and distributed the reverse-engineered distribution among it evenly. Having fit each regression, we then computed the BIC for each regression from the residual sum of squares, which we derived from the segmented output, the number of parameters given in the regression model, and the sample size, which was computed using the given sample sizes per age bracket of each test. For each subscale we selected the regression with the lowest BIC for our final conclusions.

Results

The results of our analysis indicate that, for the three most commonly used measures of IQ, the assumption of linearity between MA, CA, and IQ is not always true and varies by subtest. That is, the WASI-II, WISC-V, and SB5 do not abide by the assumption of linearity overall, at least at the subscale level. Here, we describe general trends across the subscales of the three tests that were examined.

The most important results of the analysis are the ones presented in Table 2 and Figure 1, with supporting information in Table 3 and Figures 2–5. Table 2 reflects the relative BIC that best described the nature of the distribution of the subscales as a function of age. For reference, the BIC is a criterion for selecting one model over another among a set. Similar to correlations or effect sizes, BIC sizes are associated with different indicators of fit. The critical information in the BIC is its size relative to other BICs; therefore, in Table 2, we present the smallest BIC for each subscale as 0.0 and other BICs for the subscale as how much larger they are than the smallest one to indicate how much better one model is over the others. A difference in BIC between 0–2 suggests that the alternative models are not worth considering, while a difference in BIC between 2–6 suggest that there is some positive evidence for an alternative model, a difference in BIC between 6–10 suggests strong evidence for the alternative model, and a difference in BIC greater than 10 indicates very strong support for an alternative model (Kass & Wasserman, 1995). As can be seen in Table 2, across all subtests for each of the IQ tests examined (with the exception of the Binet NV-VSP) there was a clear best fitting model where smallest BIC was at least two less than all others. The bolded values in the table indicate the number of segments for each subscale that best reflect its distribution, with a higher value of *N* reflecting a higher number of times the scale changes in its linearity. Figure 1 provides a visual representation of the number and approximate age of the breakpoints for all subtests. The specific age related to each of the breakpoints can be found in Table 3.

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

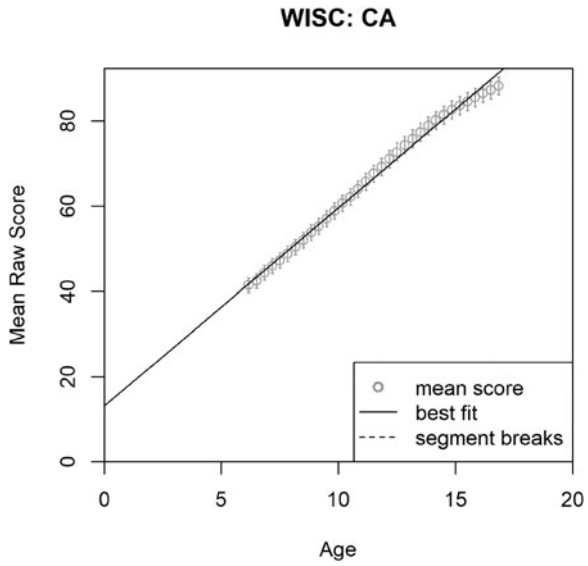

Figure 2. Representative image of linear relationship between age and raw scores (WISC-V cancellation subtest).

There were overall four different classes of models represented across the data, though the frequency of these classes were not equally distributed. Sample models and data fits for each of the four classes are presented in Figures 2–5, and each model class is described below.

In Model 1 (Figure 2), there are no changes in linearity across any age group. This would be represented in Table 2 by a BIC of zero in the first column ($N = 1$). Although this is the ideal model for MA matching, as it suggests a linear relationship between age and IQ across the age range of the test, this pattern was only noted for the WISC-V cancellation subtest (WISC CA), in which children must cross out targets presented in an array of targets and nontargets in a speeded manner.

In Model 2 (Figure 3), there is a single break point, represented in subtests where BIC is zero in the $N = 2$ column. All of the subtests of the WASI-II follow this model, though the age at which breakpoints occur between slopes differ for each subtest (see Table 3). For example, the best fit model for the WASI-II matrix reasoning (WASI MR) subtest (row 3 of Table 2) has two segments (BIC = 0 at $N = 2$), suggesting that two regression lines of different slopes best explain the data. Looking at the corresponding row in Table 3, one can see that this breakpoint occurs just before age 10, suggesting that the relationship between raw score and age changes at this age on this subtest. By looking at Figure 1, one can examine the age at which these breakpoints differ relative to other subtests specifically.

Models 3 and 4 have multiple break points in each subtest, but these occur for very different reasons. In Model 3 (Figure 4), which represent the WISC-V symbol search (WISC SS) and coding (WISC CD) subscales, there is a stark break point and change in slope that begins again at floor. This pattern reflects the fact that the test itself changes between those age brackets. For example, as can be seen in Figures 2, which is based on the data for the symbol search subtest, there is an abrupt change at age 8. On this subtest children 6–7 years of age have 2 minutes (120 seconds) to determine whether a target symbol presented on the left of the page is also present in a group of five symbols on the right side of the page. Children are instructed to mark a box labelled "No" if the search group does not contain the target symbol on the left. In contrast, for children between the ages of 8 and 16 years, there are two target symbols, rather than one, and children mark "No" if neither of target symbols are present. Thus, raw scores, which reflect number of items correct begin again at zero at the age-8-year mark, and there is therefore a change in slope at 7 years, 11 months, and 30 days as, as well as one at 8 years, 0 months, and 1 day, accounting for two of the three break points in this model. In addition, there is a third break point around age 14 for the coding subtest and around 12 for the symbol search subtest that reflect a change in slope that is unrelated to a change in instruction. However, both before and after the breakpoint, the slopes reflecting changes in raw scores as a function of age are, as in Model 1, linear, suggesting that matching MAs within each set of slopes would work well, but that matching across those ranges is problematic.

Finally, Model 4 reflects multiple break points. This pattern is reflected most clearly in the Stanford–Binet subtests where there were three to four breakpoints in each subtest. For example, on the Stanford–Binet nonverbal fluid reasoning subscale, a routing subtest, there were three breakpoints (BIC = 0 at $n = 4$), with changes in slope around age 5, 7, and 11 years of age (see Figure 5). These changes are commensurate with changes in the materials used in the tests, with children under 5 being presented with manipulatives and then beginning around age 7, the materials switch so that children are now being asked to respond to standard matrices.

Discussion

MA matching has provided a methodological tool with which to improve the scientific rigor of research on intellectual disability, and a means with which to conceptualize, operationalize, and actualize Ed's developmental approach to intellectual disaibility. The use of matching strategies has allowed us to recast our understanding and focus away from deficits or defects, in favor of a more nuanced look at the relation between specific areas of performance and more general cognitive function. This methodological tool has led to more humane, compassionate, and scientifically accurate narratives about the strengths and weaknesses of individuals with specific conditions, and serves both to help us understand atypical development and to underscore the inherent universality of development. The more precise and accurate we are in our use of MA-matching strategies, the better able we will be to truly understand the developmental trajectories of those we aim to serve.

Much has been debated about what MA measures and metrics should be used to answer which types of research questions, and how comparison groups should be constructed to optimize the veracity of the knowledge to be gained (see Burack, 2004). What has become clear from this line of research is that IQ tests and the scores we derive from them, including MAs, have a broad impact. What has been less clear is information about how the tests that we choose impact our data. Here we demonstrate that rates of knowledge acquisition on IQ tests are not uniformly linear either across or within tests, making plain a need to consider how linearity changes within each subtest in our understanding of how best to match, which subtests to use at what ages, and the impact of specific matching strategies on any conclusions.

Recommendations

Overall, the data suggest that the most linear subtests are the ones we can have the most confidence in using to compare groups of

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

Table 2. Relative Bayesian information criteria (BICs) for segmented regressions for *N* segments

		Number of segments					
Test	Subscale	*N* = 1	*N* = 2	*N* = 3	*N* = 4	*N* = 5	*N* = 6
WASI	Block design	32.0	**0.0**	13.0	27.8	42.6	57.7
WASI	Vocabulary	79.8	**0.0**	11.2	25.2	40.0	55.0
WASI	Matrix reasoning	134.4	**0.0**	11.9	26.6	41.7	56.9
WASI	Similarities	41.9	**0.0**	13.2	26.7	41.8	57.0
Binet	NV-FR	2156.7	111.1	102.3	**0.0**	13.3	29.1
Binet	NV-K	965.3	159.5	**0.0**	5.9	21.0	35.5
Binet	NV-QR	1782.1	170.6	**0.0**	10.5	16.6	31.5
Binet	NV-VSP	1426.1	*0.7*	**0.0**	13.4	26.7	41.7
Binet	NV-WM	1749.4	56.6	**0.0**	11.0	17.5	31.9
Binet	V-FR	1335.0	32.3	21.8	**0.0**	13.1	24.4
Binet	V-K	951.8	27.2	**0.0**	5.3	15.5	30.8
Binet	V-QR	1231.5	8.6	**0.0**	11.8	9.3	25.3
Binet	V-VSP	1190.8	45.5	4.2	**0.0**	11.7	26.5
Binet	V-WM	1926.7	17.8	**0.0**	9.1	25.0	40.7
WISC	BD	46.1	**0.0**	14.7	29.9	45.7	61.6
WISC	SI	97.3	**0.0**	9.6	25.6	41.5	174.3
WISC	MR	171.4	**0.0**	11.6	26.1	42.1	58.0
WISC	DS	64.4	**0.0**	9.5	25.4	125.9	57.3
WISC	CD*	2.8	10.5	33.6	**0.0**	15.6	31.6
WISC	VC	58.4	**0.0**	14.2	29.6	45.2	61.2
WISC	FW	90.9	**0.0**	15.4	31.3	47.3	63.1
WISC	VP	77.0	**0.0**	12.0	27.7	43.7	59.5
WISC	PS	80.2	**0.0**	12.3	27.9	43.8	59.9
WISC	SS*	144.2	135.0	174.9	**0.0**	15.5	31.4
WISC	IN	114.4	**0.0**	13.3	28.5	42.6	58.0
WISC	PC	78.8	**0.0**	13.3	28.9	44.0	59.8
WISC	LN	125.2	**0.0**	7.9	22.4	37.9	53.4
WISC	CA	**0.0**	5.2	20.9	36.8	52.9	68.9
WISC	CO	54.8	**0.0**	14.7	30.3	45.9	61.9
WISC	AR	220.1	**0.0**	6.9	22.7	38.7	54.5

Note. Bold: zero values, smallest BIC and best fit. Italicized: BIC that is no more than 2 greater than smallest value for that subscale. *The CD and SS subscales of the WISC have three segments in their best fit, where the first segment is discontinuous with the second and third. See Figure 2 for a visualization of this.

individuals with intellectual disability to those who are typically developing across all age groups. These tests include mostly those subtests that have a timed component, such as the WISC-V cancellation, symbol search, and coding subtests. These subtests all map on to the processing speed scale of the WISC-V and might represent a reasonable matching strategy to use for experimenters who are trying to compare groups on computerized tasks in which measurements of reaction time and accuracy are central.

Matching on the basis of processing speed may not be relevant to a research group's empirical question. If this is the case, then researchers to be mindful about the range of CAs and MAs they are considering and how that relates to the subtest they are using for matching purposes. Wherever possible researchers should try to stay "within slope" on the matching subtest. For example, the knowledge subtest of the SB5 shows changes in slopes at aroud 5 and 10 years of age. If matching on this subtest is relevant to the particular research question being asked, then it would be adviseable to test children between CA and MAs between 5 and 10 years (where there is a linear relationship between CA and IQ) rather then testing children whose CA and or MAs straddle a break (e.g. 7–12-year-olds). While this might seem constraining, it provides some assurance that the matching strategy that is adopted by the researcher also fits with the fundamental structure of the test being used to match.

The impact of these discontinuities in linearity can have considerable impact on researchers' interpretation of research

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

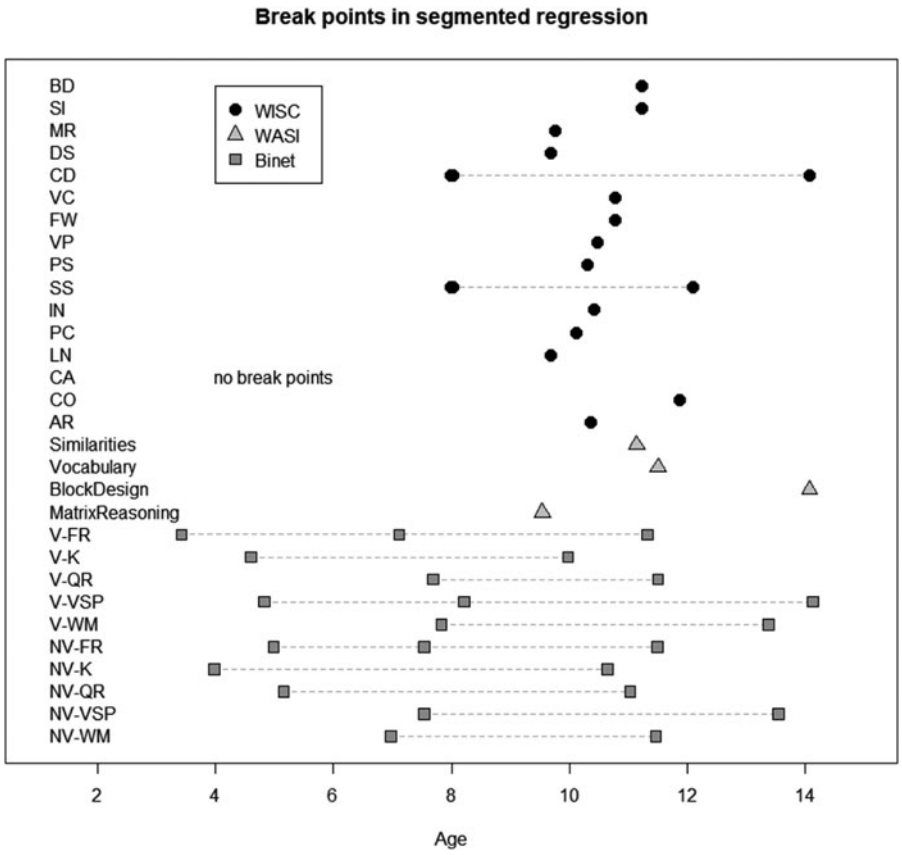

Figure 1. Breakpoints of each subtest of each intelligence scale.
Note. WISC-V subtest order is organized as a function of whether they are primary of secondary subtests with abbreviations representing the following subtests: AR = arithmetic; BD = block design; CA = cancellation; CD = coding; CO = comprehension; DS = digit span; FW = figure weights; IN = information; LN = letter–number sequences; MR = matrix reasoning; PC = picture concepts; PS = picture span; SI = similarities; SS = symbol search; VC = vocabulary; VP = visual puzzles. WASI-II subtests are grouped by verbal and then nonverbal tasks. For the SB-5, the verbal subtests are presented first followed by the nonverbal subtest and the abbreviations are as follows: V- = verbal; NV- = nonverbal; FR = fluid reasoning; K = knowledge; QR = quantitative reasoning; VSP = visual-spatial processing; WM = working memory.

findings. This applies both to prospective scientific inquiry, as well as to interpreting currently published literatures. This increased precision and nuance in our understanding and our measurements will allow for a clear developmental picture of the strengths and weaknesses of those with intellectual disaibility, within the context of a developmental approach, to emerge.

Limitations

One relevant critique of the analyses presented here is that MA matching is typically conducted at the level of either FSIQ or some other IQ composite, for example Verbal IQ (VIQ) or Performance IQ (PIQ), rather than at the subscale level. The primary reason for conducting these analyses at the subscale level is that composite scores, such as FSIQ, are based on the composite of standardized scores of various subtests. For instance, the VIQ composite score of the WASI-II represents a conversion of the sum of the vocabulary and similarities subtest standardized scores. The conversion from standardized score sums to composite score does not include age as a variable because it has already been accounted for in the conversion from raw score to standardized score. In other words, the linearity assumptions we set out to evaluate could only be tested at the subscale level. The underlying structures of these subscales are a necessary foundation to understand how they are combined to create our conceptualization of these latent constructs. The farther away our measures get from the construct of interest, the greater the likelihood that our assumptions will not be met.

The mathematical evidence we provide in our analyses appear to reflect and support widely held theories of typical child cognitive development. In general, slopes at younger ages are significantly steeper than those at the older ages, highlighting that the rate of cognitive development at these ages is faster than at older ones. This is particularly evident in the finding of a breakpoint at age 5 years for most of the subtests of the Stanford–Binet that is contrasted with the fact that both the WASI-II and the WISC-V can only first be administered at 6 years of age. These data help provide empirical support for our understanding that IQ scores tend to be unstable in early childhood: cognitive functioning is rapidly developing during the early childhood years and the steeps slopes presented in Figures 2–5 imply that small variations in raw scores can result in larger changes in standard scores at younger ages than they do at older ages. Examining developmental trends in later childhood, 23 out of the 29 subtests evaluated in our analyses demonstrated a break between the ages of 9.5 to 12.0 years old. These breaks represent a "flattening" of the slope representing the rate at which raw score changes as a function of age. This shift maps on to the preadolescent period, or the transition from concrete operational to formal operational stages in Piaget's theory of development (Piaget, 1972).

An important limitation of the present analyses is that they reflect norms collected on large datasets of typically developing children and adolescents. While this type of norming work with typically developing populations is intensive and important, it may be less relevant to our understanding of cognitive developmental trajectories of children and adolescents with intellectual disability, especially if associated with a specific genetic syndrome. As the rate of children diagnosed with intellectual disability continues to increase (Zablotsky et al., 2019), so too does the pressing need for test developers to create validated testing norms for these children. These norms are crucial in order to provide clinicians

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

Table 3. Ages (in years) at each breakpoint for each IQ subtest

		Age (in years)		
Test	Subscale	Break 1	Break 2	Break 3
WISC	BD	11.23		
WISC	SI	11.23		
WISC	MR	9.76		
WISC	DS	9.68		
WISC	CD*	7.98	8.02	14.06
WISC	VC	10.77		
WISC	FW	10.78		
WISC	VP	10.47		
WISC	PS	10.30		
WISC	SS*	7.98	8.02	12.10
WISC	IN	10.41		
WISC	PC	10.12		
WISC	LN	9.69		
WISC	CA	0.00	0.00	0.00
WISC	CO	11.87		
WISC	AR	10.36		
WASI	Block design	14.07		
WASI	Vocabulary	11.50		
WASI	Matrix reasoning	9.54		
WASI	Similarities	11.14		
Binet	NV-FR	4.98	7.53	11.49
Binet	NV-K	3.98	10.64	
Binet	NV-QR	5.15	11.03	
Binet	NV-VSP	7.53	13.54	
Binet	NV-WM	6.97	11.46	
Binet	V-FR	3.42	7.11	11.32
Binet	V-K	4.60	9.97	
Binet	V-QR	7.69	11.50	
Binet	V-VSP	4.82	8.21	14.13
Binet	V-WM	7.83	13.37	

AR = arithmetic; BD = block design; CA = cancellation; CD = coding; CO = comprehension; DS = digit span; FW = figure weights; IN = information; LN = letter–number sequences; MR = matrix reasoning; PC = picture concepts; PS = picture span; SI = similarities; SS = symbol search; VC = vocabulary; VP = visual puzzles. WASI-II subtests are grouped by verbal and then nonverbal tasks. For the SB-5, the verbal subtests are presented first followed by the nonverbal subtest and the abbreviations are as follows: FR = fluid reasoning; K = knowledge; V- = verbal; NV- = nonverbal; QR = quantitative reasoning; VSP = visual-spatial processing; WM = working memory

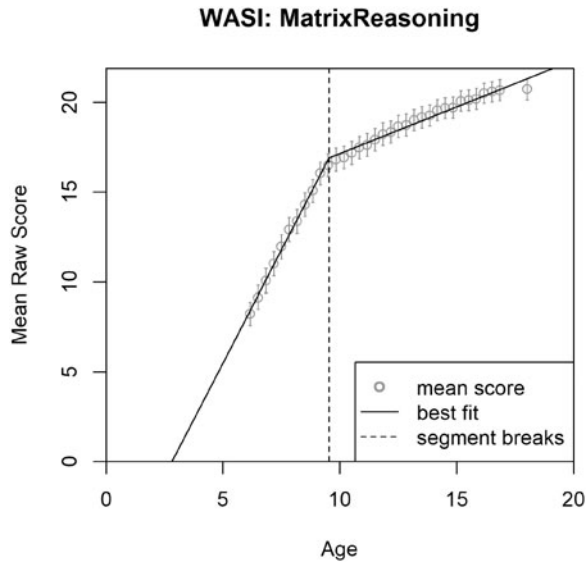

Figure 3. Representative image of distribution with two regression lines to fit relationship between age and raw scores (WASI-II matrix reasoning subtest).

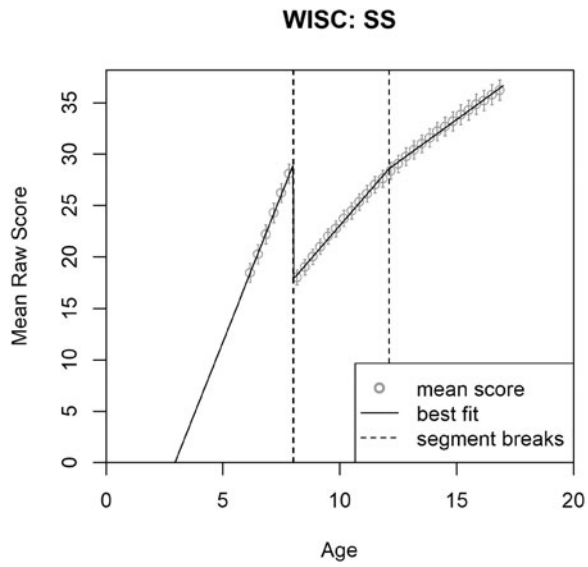

Figure 4. Representative image of distribution with three regression lines to fit relationship between age and raw scores (WISC-V symbol search).

and stakeholders with information regarding their child(ren)'s development.

As our analyses demonstrate, even with widely used and generally supported tools, cognitive development cannot be captured in a single test score, or even a combination of multiple subtest scores. Heeding the ever-relevant lessons of Ed Zigler, we also recognize that cognitive development is but a small component of children's overall development. Perhaps, we, as researchers, focus on cognitive development because we have tools in which we have invested decades of intellectual time and effort and, as such, also our confidence in what the measures represent. However, as researchers, we are obliged to interpret such unidimensional snapshots of development as exactly that. While we hope that the unpacking of the concept of MA and providing a pragmatic examination of the limits of its measurement will advance the empirical practice of MA matching in future research, we recognize that it is not a panacea for developmental science. We use these tools to make comparisons between groups of individuals and draw conclusions from these data about development; however, development cannot be adequately captured in a single time point.

Relevant to both clinical and research practices, the original authors of the IQ tests examined here, Binet and Weschler, each acknowledged the limits of their tests. Binet noted that IQ

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

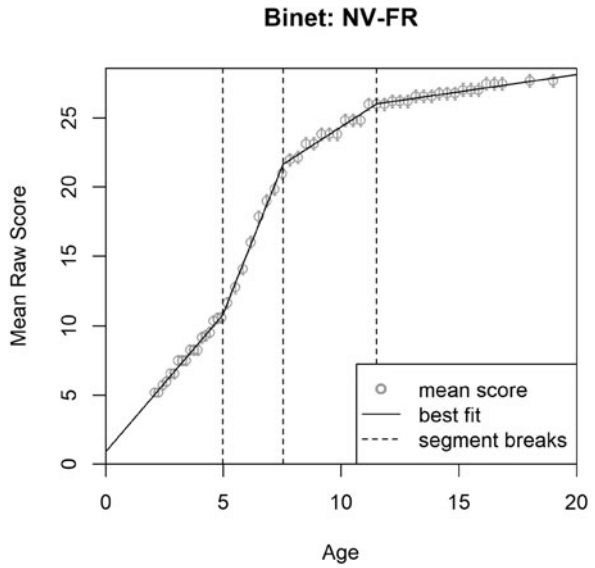

Figure 5. Representative image of distribution with four regression lines to fit relationship between age and raw scores (SB5 nonverbal – fluid reasoning subtest).

testing should only be one part of the complete assessment of an individual, whereas Wechsler spent 30 years of his career attempting to develop tests of what he called "nonintellective factors" such as drive, motivation, and persistence (among others) that impacted an individual's performance and would "correlate sufficiently with the full-scale scores, and yet emerge as factorially different" (Wechsler, 1981, p. 85). That is, in spite of their monumental bodies of work and the immeasurable influence the measures these two individuals developed, they both knew that there was more to "intelligence," broadly defined than could be measured on a single test. Zigler also clearly agreed with this in his quest to understand and work with the whole person. "Cognitive skills are very important, but they are so intertwined with the physical, social, and emotional systems that it is shortsighted, if not futile, to dwell on the intellect and exclude its partners" (Zigler & Bishop-Josef, 2006, p. 22).

Acknowledgment. The authors would like to acknowledge the families of children with disabilities who continue to entrust us with the task of furthering our understanding of development through their participation in research.

Financial Statement. This work was supported by the National Institutes of Health (MH101536-01).

Conflicts of Interest. None

References

Abbeduto, L., Thurman, A. J., Bullard, L., Nelson, S., & McDuffie, A. (2019). Genetic syndromes associated with intellectual disabilities. In C Armstrong, & L Morrow (Eds.), *Handbook of medical neuropsychology* (pp. 263–299). Cham: Springer.

Bellugi, U., Bihrle, A., Jernigan, T., Trauner, D., & Doherty, S. (1990). Neuropsychological, neurological, and neuroanatomical profile of Williams syndrome. *American Journal of Medical Genetics*, *125*(Suppl. 6), 115–125. doi:10.1002/ajmg.1320370621

Bennett-Gates, D., & Zigler, E. (1998). Resolving the developmental-difference debate: An evaluation of the triarchic and systems theory models. In J.A. Burack, R. M. Hodapp, & E. Zigler (Eds.), *Handbook of mental retardation and development* (pp. 115–131). Cambridge, UK: Cambridge University Press.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the observed learning outcome)*. New York: Academic Press.

Binet, A. (1911). New investigations upon the measure of the intellectual level among school children. In H. H. Goddard (Ed.), *Development of intelligence in children (the Binet-Simon Scale)* (pp. 274–328). Baltimore: Williams & Wilkins.

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, *24*, 383–405. doi:10.1076/jcen.24.3.383.981

Burack, J. A. (1997). The study of atypical and typical populations in developmental psychopathology: The quest for a common science. In S. S. Luthar, J. A. Burack, D. Cicchetti & J. R. Weisz (Eds.), *Developmental psychopathology: Perspectives on adjustment, risk and disorder* (pp. 139–165). New York, NY: Cambridge University Press.

Burack, J. A. (2004). Editorial preface. *Journal of Autism and Developmental Disorders*, *34*, 3–5.

Burack, J. A., Cohene, K., & Flores, H. (2011). Developmental models as frameworks for early intervention with children with Down syndrome. In J.-A. Rondal, J. Perera, & D. Spiker (Eds.), *Biocognitive rehabilitation of Down syndrome: The early years* (pp. 142–152). New York: Cambridge University Press. doi: 10.12970/2310-8231.2014.02.03.5.

Burack, J. A., Dawkins, T., Stewart, J., Iarocci, G., & Russo, N. (2012a). The mysterious myth of attention deficit…" revisited: A discussion of how the developmental approach is transforming the understanding of intellectual disability. *International Review of Research in Developmental Disabilities*, *42*, 147–177.

Burack, J. A., Evans, D. W., Klaiman, C., & Iarocci, G. (2001). The mysterious myth of attention deficits and other defect stories: Contemporary issues in the developmental approach to mental retardation. *International Review of Research in Mental Retardation*, *24*, 299–320. doi:10.1016/s0074-7750(01)80012-4

Burack, J. A., Iarocci, G., Bowler, D., & Mottron, L. (2002). Benefits and pitfalls in the merging of disciplines: The example of developmental psychopathology and the study of persons with autism. *Development and Psychopathology*, *14*, 225–237. doi:10.1017/S095457940200202X

Burack, J. A., Iarocci, G., Flanagan, T. D., & Bowler, D. M. (2004). On mosaics and melting pots: Conceptual considerations of comparison and matching strategies. *Journal of Autism and Developmental Disorders*, *34*, 65–73. doi:10.1023/B:JADD.0000018076.90715.00

Burack, J. A., Russo, N., Flores, H., Iarocci, G., & Zigler, E. (2012b). The more you know the less you know, but that's OK: Developments in the developmental approach to intellectual disability. In J. A. Burack, R. M. Hodapp, G. Iarocci & E. Zigler (Eds.), *The Oxford handbook of intellectual disability and development* (pp. 1–15). New York: Oxford University Press. doi:10.1093/oxfordhb/9780195305012.013.0001

Burack, J. A., Russo, N. N., Gordon Green, C., Landry, O., & Iarocci, G. (2016b). Developments in the developmental approach to intellectual disability. In D. Cicchetti (Ed.), *Handbook on developmental psychopathology* (3rd ed., pp. 1–55). New York: Wiley.

Burack, J. A., Russo, N., Kovshoff, H., Palma Fernandes, T., Ringo, J., Landry, O., & Iarocci, G. (2016a). How I attend – not how well do I attend: Rethinking developmental frameworks of attention and cognition in autism spectrum disorder and typical development. *Journal of Cognition and Development*, *17*, 553–567. doi:10.1080/15248372.2016.1197226

Campbell, C., Landry, O., Russo, N., Flores, H., Jacques, S., & Burack, J. A. (2013). Cognitive flexibility among individuals with Down syndrome: Assessing the influence of verbal and nonverbal abilities. *American Journal on Intellectual and Developmental Disabilities*, *118*, 193–200. doi:10.1352/1944-7558-118.3.193

Case, R. (1980). The underlying mechanism of intellectual development. In J. Kirby, & J. B. Biggs (Eds.), *Cogniton, development, and instruction* (pp. 1–38). New York: Academic Press.

Chiurazzi, P., & Pirozzi, F. (2016). Advances in understanding - genetic basis of intellectual disability. *F1000Research*, *5*, 1–16. doi:10.12688/f1000research.7134.1

Cicchetti, D., & Beeghly, M. (Eds.). (1990). *Children with Down syndrome: A developmental perspective*. New York: Cambridge University Press.

https://doi.org/10.1017/S0954579420001480 Published online by Cambridge University Press

Cicchetti, D., & Ganiban, J. (1990). The organization and coherence of developmental processes in infants and children with Down syndrome. In R. M. Hodapp, J. A. Burack & E. Zigler (Eds.), *Issues in the developmental approach to mental retardation mental retardation* (pp. 169–225). New York: Cambridge University Press.

Cicchetti, D., & Pogge-Hesse, P. (1982). Possible contributions of the study of organically retarded persons to developmental theory: The developmental-difference controversy. In E. Zigler, & D. A. Balla (Eds.), *Mental retardation: The developmental-difference controversy* (pp. 277–318). New York: Lawrence Erlbaum Associates.

Flanagan, T., Russo, N., Flores, H., & Burack, J. A. (2008). The developmental approach to the study of down syndrome: Conteporary issues in historical perspective. *Down Syndrome Research and Practice*, *Online*, 96–100.

Hanaoka, T., Mita, K., Hiramoto, A., Suzuki, Y., Maruyama, S., Nakadate, T., … Egusa, Y. (2010). Survival prognosis of Japanese with severe motor and intellectual disabilities living in public and private institutions between 1961 and 2003. *Journal of Epidemiology*, *20*, 77–81. doi:10.2188/jea.JE20090024

Hodapp, R. M., Burack, J. A., & Zigler, E. (1990). The developmental perspective in the field of mental retardation. In R. M. Hodapp, J. A. Burack, & E. Zigler (Eds.), *Issues in the developmental approach to mental retardation* (pp. 3–26). New York: Cambridge University Press.

Hodapp, R. M., & Zigler, E. (1995). Past, present, and future issues in the developmental approach to mental retardation and developmental disabilities. *Developmental Psychopathology*, *2*, 299–331.

Hoekstra, R. A., Bartels, M., & Boomsma, D. I. (2007). Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. *Learning and Individual Differences*, *17*, 97–114. doi:10.1016/j.lindif.2007.05.005

Iarocci, G., & Burack, J. A. (1998). Understanding the development of attention in persons with mental retardation: Challenging the myths. In J. A. Burack, R. M. Hodapp, & E. Zigler (Eds.) *Handbook of mental retardation and development* (pp. 349–381). New York: Cambridge University Press.

Iwase, S., Bérubé, N. G., Zhou, Z., Kasri, N. N., Battaglioli, E., Scandaglia, M., & Barco, A. (2017). Epigenetic etiology of intellectual disability. *Journal of Neuroscience*, *37*, 10773–10782. doi:10.1523/JNEUROSCI.1840-17.2017

Jarrold, C., & Brock, J. (2004). To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders*, *34*, 81–86. doi:10.1023/B:JADD.0000018078.82542.ab

Karam, S. M., Barros, A. J. D., Matijasevich, A., Dos Santos, I. S., Anselmi, L., Barros, F., … Black, M. M. (2016). Intellectual disability in a birth cohort: Prevalence, etiology, and determinants at the age of 4 years. *Public Health Genomics*, *19*, 290–297. doi:10.1159/000448912

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, *90*, 928–934. doi:10.1080/01621459.1995.10476592

Katz, G., & Lazcano-Ponce, E. (2008). Intellectual disability: Definition, etiological factors, classification, diagnosis, treatment and prognosis. *Salud Pública de México*, *50*, s132–s141. Retrieved from https://www.medigraphic.com/pdfs/salpubmex/sal-2008/sals082e.pdf%0Ahttp://www.scielosp.org/scielo.php?script=sci%7B_%7Darttext%7B&%7Dpid=S0036-36342008000800005%7B&%7Dlang=pt

Koenen, K. C., Moffitt, T. E., Roberts, A. L., Martin, L. T., Kubzansky, L., Harrington, H., … Caspi, A. (2009). Childhood IQ and adult mental disorders: A test of the cognitive reserve hypothesis. *American Journal of Psychiatry*, *166*, 50–57. doi:10.1176/appi.ajp.2008.08030343

Lane, K. A., Stewart, J., Fernandes, T., Russo, N., Enns, J. T., & Burack, J. A. (2014). Complexities in understanding attentional functioning among children with fetal alcohol spectrum disorder. *Frontiers in Human Neuroscience*, *8*(March), 1–8. doi:10.3389/fnhum.2014.00119

McCall, R. B. (1977). Childhood IQ's as predictors of adult educational and occupational status. *Science*, *197*, 482–483.

Mervis, C.B., & Robinson, B.F. (1999). Methodological issues in cross-syndrome comparisons: Matching procedures, sensitivity (Se), and specificity (Sp). *Monographs of the Society for Research in Child Development*, *64*, 115–130.

Mervis, C. B., & Klein-Tasman, B. P. (2000). Williams syndrome: Cognition, personality, and adaptive behavior. *Mental Retardation and Developmental Disabilities Research Reviews*, *6*, 148–158.

Mervis, C. B., & Klein-Tasman, B. P. (2004). Methodological issues in group-matching designs: α levels for control variable comparisons and measurement characteristics of control and target variables. *Journal of Autism and Developmental Disorders*, *34*, 7–17. doi:10.1023/B:JADD.0000018069.69562.b8

Mir, Y. R., & Kuchay, R. A. H. (2019). Advances in identification of genes involved in autosomal recessive intellectual disability: A brief review. *Journal of Medical Genetics*, *56*, 567–573. doi:10.1136/jmedgenet-2018-105821

Muggeo, V. M. R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine*, *22*, 3055–3071.

Muggeo, V. M. R. (2008). Segmented: An R package to fit regression models with broken-line relationships. *R News*, *8*, 20–25.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development*, *15*, 1–12.

Roid, G. H. (2003). *Stanford-Binet intelligence scales, fifth edition, technical manual*. Itasca, IL: Riverside Publishing.

Russo, N., Flanagan, T., Iarocci, G., Berringer, D., Zelazo, P. D., & Burack, J. A. (2007). Deconstructing executive deficits among persons with autism: Implications for cognitive neuroscience. *Brain and Cognition*, *65*, 77–86. doi:10.1016/j.bandc.2006.04.007

Schneider, W., Niklas, F., & Schmiedeler, S. (2014). Intellectual development from early childhood to early adulthood: The impact of early IQ differences on stability and change over time. *Learning and Individual Differences*, *32*, 156–162. doi:10.1016/j.lindif.2014.02.001

Selman, R. L. (1980). *The growth of interpersonal understanding: Developmental and clinical analyses*. New York: Academic Press.

Stevenson, R. E. (2000). Splitting and lumping in the nosology of XLMR. *American Journal of Medical Genetics –Seminars in Medical Genetics*, *97*, 174–182. doi:10.1002/1096-8628(200023)97:3<174::AID-AJMG1034>3.0.CO;2-4

Vissers, L. E. L. M., Gilissen, C., & Veltman, J. A. (2016). Genetic studies in intellectual disability and related disorders. *Nature Reviews Genetics*, *17*, 9–18. doi:10.1038/nrg3999

Vorstman, J. A. S., & Ophoff, R. A. (2013). Genetic causes of developmental disorders. *Current Opinion in Neurology*, *26*, 1063–1078. doi:10.1007/s00213-013-3334-z

Wechsler, D. (1932). Analytic use of the Army Alpha. *Journal of Applied Psychology*, *16*, 254–256. doi:10.1037/h0072688

Wechsler, D. (1981). The psychometric tradition: Developing the Wechsler adult intelligence scale. *Contemporary Educational Psychology*, *6*, 82–85. doi:10.1016/0361-476X(81)90035-7

Wechsler, D. (2011). *Wechsler abbreviated scale of intelligence, second edition (WASI-II): Manual*. San Antonio, TX: NCS Pearson.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children –fifth edition (WISC-V): Technical and interpretive manual*. Bloomington, MN: Pearson.

Wolfe, K., Strydom, A., & Bass, N. (2019). Genetics of intellectual disability. In M. Scheepers, & M. Kerr (Eds.), *Seminars in the psychiatry of intellectual disability* (pp. 12–27). New York: Cambridge University Press.

Zablotsky, B., Black, L. I., Maenner, M. J., Schieve, L. A., Danielson, M. L., Bitsko, R. H., … Boyle, C. A. (2019). Prevalence and trends of developmental disabilities among children in the United States: 2009–2017. *Pediatrics*, *144*, 2009–2017. doi:10.1542/peds.2019-0811

Zigler, E. (1967). Familial mental retardation: A continuing dilemma. *Science*, *155*, 292–298. Retrieved from https://www.jstor.org/stable/1720653

Zigler, E. (1969). Developmental versus difference theories of mental retardation and the problem of motivation. *American Journal of Mental Deficiency*, *73*, 536–556.

Zigler, E., & Balla, D. A. (Eds.). (1982). *Mental retardation: The developmental-difference controversy*. Hillsdale, NJ: Lawrence Erlbaum.

Zigler, E., & Bishop-Josef, S. J. (2006). The cognitive child versus the whole child: Lessons from 40 years of head start. In D. G. Singer, R. M. Golinkoff, & K. Hirsh-Pasek (Eds.), *Play=Learning: How play motivates and enhances children's cognitive and social-emotional growth* (pp. 15–35). New York: Oxford University Press. doi:10.1093/acprof.

Zigler, E., & Hodapp, R. M. (1986). *Understanding mental retardation*. New York, NY: Cambridge University Press.