International Journal Educational Psychology
ΨJEP

Hipatia Press
www.hipatiapress.com

Instructions for authors, subscriptions, and further details:
http://ijep.hipatiapress.com/

Quality of Child Development Scales. A Systematic Review

Sara M. Luque de Dios[1], Araceli Sánchez-Raya[1], Juan A. Moriana[1]

[1]University of Córdoba, Spain

Date of publication: June 24th, 2023
Edition period: June 2023-October 2023

To cite this article: Luque de Dios, S.M., Sánchez-Raya, A., & Moriana, J. A. (2023). Quality of Child Development Scales. A Systematic Review. *International Journal of Educational Psychology, 12*(1) 92-124.
doi: http://doi.org/10.17583/ijep.10773

To link this article: http://dx.doi.org/10.17583/ijep.10773

IJEP – International Journal of Educational Psychology Vol. 12 No.2-Online First

Quality of Child Development Scales. A Systematic Review

Sara M. Luque de Dios
University of Córdoba
Juan A. Moriana
University of Córdoba
Araceli Sánchez-Raya
University of Córdoba

Abstract

Developmental scales for children aged 0-6 years are a particularly valuable resource for assessing developmental milestones in children. Most scales are developed based on a broad conceptual framework, and their metric validation is insufficient and of low quality. The aim of this systematic review is to analyse the psychometric quality of these tests and identify aspects in need of improvement. To this end, the PRISMA methodology and the WOS and ProQuest databases were used to search for articles addressing this topic. A total of 680 articles were identified, of which 72 were selected using the established inclusion and exclusion criteria. The results indicate a scarcity of independent studies on the statistical measurement of the scales. The selected articles are very heterogeneous and validate these tests using adaptations of common metrics. Most perform cross-cultural, concurrent, and prognostic validations of the tests. We conclude that the quality of the scale metrics and other common aspects of these tests need to be improved, particularly sample sparsity and heterogeneity, as well as cultural biases. We underline the importance of applying for advances in metrics for the construction of developmental scales and recommend the use of computerised versions to improve their ease of use and efficiency.

Keywords: developmental scales, psychometrics, systematic review, assessment instruments, evolutionary development.

ISSN: 2014-3591
DOI: 10.17583/ijep.10773

Hipatia Press
www.hipatiapress.com

IJEP – International Journal of Educational Psychology Vol. 12 No.2-Online First

Calidad de las Escalas del Desarrollo Infantil. Una Revisión Sistemática

Sara M. Luque de Dios
Universidad de Córdoba
Araceli Sánchez-Raya
Universidad de Córdoba
Juan A. Moriana
Universidad de Córdoba

Resumen

Las escalas de evaluación del desarrollo destinadas a menores de 0 a 6 años son un recurso muy importante para valorar los hitos evolutivos de la población infantil. La mayoría de ellas presentan un marco conceptual amplio y su validación métrica es insuficiente y de baja calidad. El objetivo de esta revisión sistemática es analizar la calidad psicométrica de estas pruebas y señalar aspectos susceptibles de mejora. Se ha seguido la metodología PRISMA y las bases de datos WOS y PROQUEST, encontrando un total de 680 artículos, seleccionando finalmente 72 documentos relacionados, una vez aplicados los criterios de inclusión y exclusión. Los resultados muestran un escaso número de estudios independientes dedicados a la medición estadística de las escalas. Los trabajos encontrados son muy heterogéneos y aplican a estas pruebas adaptaciones de la métrica común para su validación. La mayoría de los artículos realizan validaciones transculturales, concurrentes y pronósticas de las pruebas. Concluimos afirmando que es necesario mejorar la calidad métrica de las escalas, señalando aspectos comunes de los que adolecen: escasez y heterogeneidad de las muestras, además de sesgos culturales. Se subraya la importancia de aplicar avances métricos en la elaboración de escalas del desarrollo y se recomienda apostar por versiones computarizadas que las hagan más cómodas y eficientes y aumenten su usabilidad.

Palabras clave: escalas de desarrollo, psicometría, revisión sistemática, instrumentos de evaluación, desarrollo evolutivo.

ISSN: 2014-3591
DOI: 10.17583/ijep.10773

Hipatia Press
www.hipatiapress.com

In early childhood, the level of development is mainly assessed using a combination of semi-structured interviews, informal observation, and the direct or indirect administration of developmental scales, the latter of which have a decisive weight in the diagnosis and intervention of children (Committee on Children with Disabilities, 2001). Developmental scales are standardised instruments that apply normative values to interpret children's scores. These tests provide information about the developmental characteristics and evolution of children in various domains and enable comparing different population subgroups, determining needs and services, planning interventions, monitoring developmental changes, and assessing the effectiveness of treatments. One of the most relevant aspects of scaled tests is their use in healthcare, education, and research. The scales are used in a wide range of populations and for a variety of conditions: different age groups, minority nationalities and ethnicities, developmental delays in different domains, and developmental disorders, rare diseases, and sensory deficits (Karasik y Robinson, 2022). Therefore, these instruments must have specific validations and standards of application (Gleason, 2010) and there should be a large body of independent papers published for each type of sample.

Developmental scales and tests are often administered in different settings, such as physician's offices, schools, child psychology centres, and even in the child's home. They should be administered in a calm environment and the child should feel a sense of control (World Health Organization, 2012). It is essential to create rapport between the child and the examiner, which is why parents are often asked to help (Adolph & Hoch, 2019). To ensure these conditions, more time may be needed to administer the tests, after which healthcare practitioners such as doctors, psychologists, speech therapists, and physiotherapists draw conclusions.

Although there is no consensus regarding the theoretical framework on which developmental scales should be built (McCoy, 2022), they are generally based on the concept of developmental milestones. Developmental milestones are observable behaviours exhibited by children that appear in a sequential manner at established stages of development. Variations in the rate or manner

of their occurrence may be indicative of immaturity or neurological disorders, but not in all cases. For example, children stand upright at around 12 months on average, but some children first stand at 10 months and others at 16 months, which does not necessarily indicate a developmental problem. Such differences could be interpreted as a sign of risk to watch out for to see how the child evolves or they may simply be resolved at a later developmental stage (Boonzaaijer et al., 2020). The most representative milestones for each age range and developmental domain are noted and converted into items. These items form the scales.

Traditionally, these scales have been validated under the classical test theory (CTT), which assumes that an individual's empirical score on a test is composed of their true score and a measurement error that is estimated by means of a linear model (Muñiz, 2010). Some scales, such as the Bayley Scales of Infant Development, Third Edition (Bayley-III; Bayley, 2006) and the Battelle Developmental Inventory, Second Edition (BDI-2; Newborg, 2005), use the basic statistical procedures of CTT and heterogeneous, insufficient, and non-representative samples.

The advances following the development of various psychometric theories, such as item response theory (IRT) and IRT models, have been scarcely implemented in this field. IRT uses probabilistic models to calculate subjects' trait level and relate it to the properties of the items on a test (Lalor & Rodriguez, 2022). The revised Merrill-Palmer Scale (MP-R; Roid, 2004) constitutes a psychometric advance in this field by applying IRT to its validation, thus overcoming some inaccuracies of earlier scales, albeit with statistical adjustments.

Literature reviews that address these aspects of developmental scale metrics are scarce. Visser et al. (2012) assessed the applicability of different scales in children with functional diversity. They concluded that the quality of the instruments needs to be improved, especially in children under 2 years of age or those with motor impairment, and that there are no suitable instruments for children with visual impairment or visual disability. Silva et al. (2018) evaluated and established an independent classification of

multidimensional scales and gave the Bayley-III, BDI-2, and Vineland-II the highest score in validity and reliability. The authors also noted that the most widely used instruments and those of highest metric quality have not been validated for developing countries (Olusanya et al., 2021). Both reviews underlined the need for research on the construction and validation of developmental scales.

Objective

The aim of this review is to assess the main multidimensional developmental scales for children aged 0 to 90 months through an analysis of statistical studies on their psychometric quality and potential limitations and strengths.

Method

A systematic review of the scientific and grey literature was carried out following the guidelines of the PRISMA statement (Page et al., 2021). The research question was formulated according to the PICO search strategy as follows: "What is the psychometric quality of the most commonly used multidimensional developmental scales?" The different phases of the review process are described in detail below.

Initial search

The first searches were conducted from January to March 2021 with the terms "develop* scale validation" and "child* develop* assess*" using the Boolean operator "AND". The databases used were the WOS which includes Core Collection, MEDLINE and SciELO, and ProQuest, which includes PsycInfo, PsycArticles, PsycBooks, and ProQuest psychology journals. This initial search provided an overview of the heterogeneity of procedures and scales.

Selection of the scales

The scales selected for their degree of clinical and research applicability and for the number of independent studies that have been published were: Merril Palmer- R, Bayley-III; Battelle-2, Brunet Lèzine Revised: Early Childhood Psychomotor Development Scale (BL-R; Josse, 1997), Pediatric Evaluation of Disabiblity Inventory (PEDI; Haley, 1992), Brazelton Neonatal Behavior Assessment Scale (NBAS; Brazelton, 1997), Child Neuropsychological Battery Second Edition- NEPSY-II (NEPSY-II; Korkman, 2007), Vineland Adaptive Behavior Assessment Scale Second Edition (Vineland-II; Sparrow et al., 2005), and Leiter International Manipulative Scale Revised (Leiter-R, Roid, 1997) (Table 1).

All these instruments are divided into developmental subdomains to assess basic processes such as cognition, attention, memory, language, motor skills, and adaptive-social behaviours. The Pediatric Evaluation of Disability Inventory (PEDI) and Vineland-II scales are an exception because they primarily assess adaptive behaviour.

Table 1

Operational characteristics of the selected instruments

Instrument	Age	Dimensions	Management
Merril Palmer- R (2004)	1-78 months	• Cognitive • Language and communication • Motor • Emotional partner • Adaptive behavior	**Direct and indirect** (30- 90 min.)
Bayley-III (2006)	1–42 months		
Battelle-2 (2005)	0-95 months		
BL-R (1997)	0-30 months	• Postural development • Manual eye coordination • Language • Social relationships • Adaptive	**Direct** (20- 45 min)

Table 1

Operational characteristics of the selected instruments (Continuation)

Instrument	Age	Dimensions	Management
NBAS (1997)	0-2 months	• Autonomic nervous system • Motor • Habituation • Organization/regulation • Social and interactive	**Direct** (20- 45 min)
NEPSY- II (2007)	3-16 years	• Attention and executive function • Language • Memory and learning • Sensorimotor perception • Social • Visuospatial processing	**Direct** (45min- 3 hours)
Leiter- R (1997)	2-21 years	• Reasoning and visualization • Attention and memory	**Direct** (20- 60 min)
PEDI (1992)	6-90 months	• Personal care • Mobility • Social functioning	**Indirect** (20-90 min)
Vineland-II (2005)	De 0 a 90 years and 11 months	• Communication • Daily living skills • Socialization • Motricity	

Eligibility criteria for independent studies

Inclusion criteria:

1. Studies that assess the metric properties of scales with samples aged less than 90 months or include this age range in a differentiated manner from the rest of the sample.

Exclusion criteria:

1. Studies that describe the scale, but do not provide a novel and independent assessment.
2. Studies assessing the metric properties of the scale with samples older than 90 months.
3. Studies using earlier, later, or different versions of the selected scales.

Systematic search

The search began in March 2021 and was completed in February 2022. In the initial search, the search terms were the names of the scales and the keywords of the articles. At the end of the review, the research was updated to include four articles on the Bayley-III scale. Mendeley software was used for purposes of bibliographic management.

Search strategy

The selection process (Figure 1) was conducted by two researchers: the principal investigator, who selected the publications, and a second researcher that was responsible for reviewing them. Disagreements were resolved by consensus.

Phase 1: documents that were duplicated between databases were excluded (n = 90).

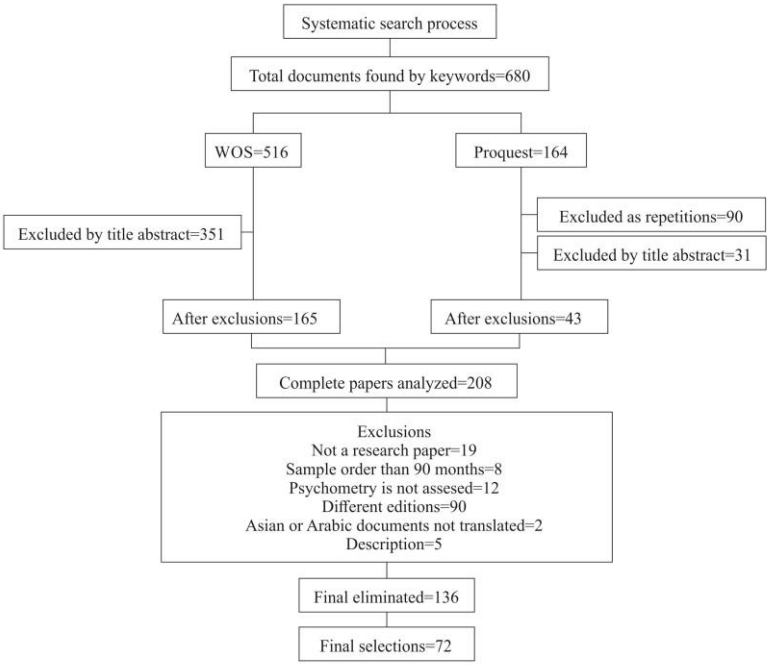
Phase 2: documents were excluded based on the information contained in the title and abstract (n = 382).

Phase 3: the full text of 208 publications was examined. Articles that included different versions of the selected scales (n = 90) as well as incomplete or missing articles (n= 19) were discarded. Articles that did not aim to make an independent assessment (n= 12) or were limited to a description of the scales (n= 5) were also discarded.

Phase 4: a total of 72 articles were finally selected: Merrill Palmer- R (n=3), Battelle II (n=4), Bayley III (n=37), Brunet Lèzine (n= 2), PEDI (n=17), NBAS (n= 4), NEPSY- II (n=1), Vineland- II (n= 2), Leiter- R (n= 2). The publications were from different countries, but all of them had been translated into English and included doctoral theses (n = 2), books (n = 1), and articles in research journals (n = 69).

Phase 5: the most important information was extracted from scientific articles and grey literature and its quality was evaluated using the following indicators: an adequate and sufficient sample, the objectives were consistent with the study, the quality and coherence of the analyses, and relevant conclusions. The risk of bias was assessed by two investigators who resolved their discrepancies by consensus.

Figure 1
PRISMA flow chart



Results

No common criteria are applied for the construction of developmental scales, except for the minimum metrics required of all types of tests. Due to this lack of uniformity, a variety of methods have been used to develop these scales.

Most scales have been validated in the framework of CTT. However, some articles validated tests for specific developmental domains using IRT with the Rasch model approach, mainly for Northern European (Berg et al., 2016), North America (Liao et al., 2004), and Asia (Yao et al., 2018) populations or adapted versions of the scales to specific populations (Amer et al., 2018) and to children with disabilities, generally older than 6 years of age (Peters, 2013).

As regards the samples used, many articles combine very broad age ranges and do not consider the differentiating characteristics of each stage of

development but establish arbitrary cut-off points by age. This occurs more frequently in samples of infants under the age of one.

Multidimensional developmental scales have not been adapted for children with disabilities or developmental disorders, although some scales have been validated independently and studies have been carried out for some representative subgroups such as children with autism spectrum disorder (ASD).

Scales designed specifically for children with functional diversity are scarce and have been little studied. Some scales have been extended to include specific disorders such as ASD or cerebral palsy (Visser et al., 2012) to the detriment of others such as sensory disabilities.

Other scales have been used for specific types of studies or population profiles regardless of whether that was their original purpose. This is the case of Vineland-II for diagnosing ASD, which has led to a subsequent increase in the number of articles published on these populations.

Most independent articles that evaluate the metric properties of scales are cross-cultural studies in populations other than those in which the scales were validated. These studies are generally conducted to adapt the tests to North American and European samples, although an increasing number of studies on Arab and Asian populations have been published.

Longitudinal studies with the same cohort using equivalent tests to evaluate the prognostic ability of the scales as well as comparative studies using different scales to measure concurrent validity have also been published.

When a scale is published or updated, it takes a significant amount of time before independent studies can be conducted or published. For this reason, it was not possible to assess newer versions of the scales.

As regards the quality of the journals where these articles were published, few appear in high impact international journals.

No results were found for local or lesser-known scales. Nor is there evidence for certain scales, such as the Haizea-Llevant Development Chart (Fernández et al., 1989) or the Carolina Curriculum (Johnson-Martin et al., 1994).

Most studies on scales that measure specific areas of development examine language skills and, to a lesser extent, motor skills. Although speech is universal, language is not, so each language or dialect requires its own validated instrument. On the other hand, measures of motor development are perhaps the most easily observable and have common characteristics that are not influenced by cultural factors.

In contrast, although cognitive development is the basis for other skills, there are fewer instruments that measure this domain, and they are more international. The socio-affective domain has only recently begun to be studied and few articles have been published, although those that have been published are very innovative.

The findings for each of the selected scales are presented below with a description of their main characteristics and factors. The analysis is intended to gain a better understanding of the scales. We begin with the most relevant scales in terms of their practical applications and the quantity and quality of independent studies on each scale. The quality of the articles is defined according to specific variables such as sample heterogeneity, the methodology and metrics used in the study, and the relevance of the conclusions.

Revised Merrill-Palmer Scale (MP-R)

The MP-R is the most recent scale for assessing development. It applies the metric advances of IRT and has shown a good fit with a quasi-random, stratified sample of 1068 children from the United States. In addition to the five subscales mentioned above, it includes three complementary scales (memory, speed of processing, and visual-motor coordination) and three indicators (social-emotion, adaptive, and self-help behaviours). The results of the MP-R can be expressed as direct scores, typical percentiles, age equivalents, and developmental scores (Rasch scores). Alcantud and Alonso (2016) compared the use of typical scores and IRT-based scores using development to determine cut-off points and found that both scoring methods are adequate.

An independent article assessed the scale in different subpopulations. Floyd et al. (2004) evaluated the cultural validity of the scale items in a

minority population of the United States and concluded that the MP-R did not exhibit differential item functioning. Peters (2013) found the scale to be sensitive for identifying developmental delay in children with ASD, but not specific for differentiating between children with ASD and children with other common disorders. The internal consistency for this sample was good but showed weak validity.

Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)

The Bayley-III has been translated and validated in the largest number of countries (Hanlon et al., 2016). A considerably larger body of independent studies has been published on the Bayley scales compared to the other scales (Table 2). This instrument has been validated with the CTT in a stratified sample of 1700 North American children. The psychometric properties of the scales are good, although they show low reliability in younger age groups (1 5 months of age), especially in receptive communication and expressive communication. The Bayley-III has been evaluated in clinical samples and the manual provides an overview of possible adaptations. It is not recommended for use with severely disabled children.

Numerous articles have compared the Bayley-III to its previous version (BSID-II) and shown that the Mental Development Index (MDI) scores are significantly higher in the new version (Moore et al., 2012; Sharp & Demauro, 2017). Correlation with the previous edition appears to be worse with lower scores (Anderson & Burnett, 2017), even for children with difficulties (Jary et al., 2013; Flynn et al., 2020). To avoid diagnostic underestimation, Lowe et al. (2012) developed an algorithm for conversion between scales.

Regarding the proposed adjustments to the scales, Milne et al. (2015) advocated averaging the ratios of the three subscales of the test for diagnosing children with functional diversity. Morsan et al. (2018) argued that gestational age correction for preterm infants should only be applied in the cognitive domain. Greene et al. (2013) showed that Bayley-III measurements of decline in average cognitive and motor skills that remain relatively stable from the

first to the second year of life in preterm infants is consistent with changes in the BSID-II. Notwithstanding, Greene et al. (2012) stated that the language indicator appears to be an important scale improvement. This domain is typically delayed in development, so they suggest the use of the sub-indexes due to the discrepancies found between receptive and expressive communication and gross and fine motor skills.

Regarding the predictive validity of the Bayley-III, correlations were found to vary for all ages and tests/subtests (Krogh & Væver, 2019a). Furthermore, this validity also varied for other scales such as the Wechsler Preschool and Primary Scale of Intelligence, Second Edition (WPPSI-II; Bode et al., 2014); the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV; Månsson et al. 2019; 2021; Nishijima et al. 2021); Peabody-2 (Lin et al., 2020); and the Movement Assessment Battery for Children, Second Edition (MABC 2; Spittle et al. 2013). Practitioners should be cautious about attributing higher Bayley-III scores to changes in direct attention. According to Krogh and Væver (2019a), predictions should be made with caution, as children at risk may be underestimated.

Mixed results have also been reported for concurrent validity between the Bayley-III and the Standford-Binet Intelligence Scale, Fifth Edition (SB-5; Kamppi & Gilmore, 2010), the McCarthy Scales of Children's Abilities (MSCA) and the Kaufman Assessment Battery for Children (KABC; Torrás Mañá et al., 2014), the Alberta Infant Motor Scale (AIMS; Alburquerque et al., 2018), and the Warner Initial Developmental Evaluation of Adaptive and Functional Skills (WIDEA-FS; Peyton at al., 2020).

Some studies have compared the normative North American sample of the Bayley-III with other populations. These studies agree on the discrepancies in scores and warn of the need to adapt the scale to different populations, even at very early ages (Vierhaus et al., 2011).

As for gender, the pattern of differences has been found to vary across scales, subscales, and ages, so it seems reasonable to assume that the Bayley III does not include gender-specific norms (Krogh & Væver, 2019b). The differences gave a higher average score to girls.

Table 2
Significant results for the Bayley-III

	Target	N	Conclusion
Lowe et al. 2012 EE.UU.	Create a conversion algorithm with the MDI of BSID-II for Bayley-III	77	High cognitive scores for Bayley-III. Creation of the conversion algorithm
Greene et al. 2012 EE.UU.	Investigate patterns and correlates of neurodevelopment	85	The language indicator is an improvement of Bayley-III
Krogh et al. 2019a Denmark	Investigate gender differences in scores	55	Differences exist, but with varying patterns.
Krogh et al. 2019b Denmark	Examine predictive validity	55	Significant correlations that varied for all ages and scales
Månsson et al. 2019 Swiss	Examine the relationship with IQ at school age	162	It is an insufficient predictor of later IQ

Battelle Developmental Inventory, Second Edition (BDI-2)

The BDI-2 is a revised edition of the original version and has been validated under the CTT framework in a stratified sample of 2,500 children from 30 US states. The reliability of this instrument is moderate to high for the total score and the different domains, but the coefficients of internal consistency are below the recommended range in several subdomains. Some articles have confirmed the psychometric robustness of the scale with the IRT Rasch model (Elbaum et al., 2010). The applicability manual indicates that the instrument has been tested in a clinical sample. It provides centile, standard, age-equivalent, T, change-sensitive, and Z scores.

As for the concurrent validity of the BDI-2, Nitsana (2010) concluded that correlations with the WPPSI-III were positive and stable. The BDI-2 is

administered to individuals with disabilities and studies have been conducted in the ASD population (Goldin et al., 2014). Sipes et al. (2011) established cut-off points for screening these children.

Pediatric Evaluation of Disability Inventory (PEDI)

The PEDI is a reference instrument for assessing physical and mental disability in children. To establish the content of the test, preliminary editions and a combination of statistical techniques, including Rasch scales and their analysis, have been evaluated. The test has been assessed with the CTT and found to present high internal consistency in a sample of 412 healthy American children. The PEDI allows calculating both scaled and standard performance scores.

The PEDI scale metrics have been assessed in independent, novel, and good quality publications and some studies have used IRT to validate this instrument (Berg et al., 2016). A computerised version with an item bank is also available (Dumas et al., 2017).

The PEDI has shown concurrent validity with the Peabody Scale as it assesses similar but not identical aspects of motor development (Mayrand et al., 2009). Results partially support the validity between the School Outcome Measure (SOM) and PEDI in pre-schoolers with functional diversity (Amundson et al., 2012).

Some studies have used the PEDI to assess individuals with disabilities, mainly cerebral palsy. Nordmark et al. (2000) and Vos-Vromans et al. (2005) compared the PEDI and the Gross Motor Function Measure (GMFM) test over time in children with cerebral palsy and concluded that both instruments were suitable and complementary as they measure different aspects of motor function.

The test should be adapted to the different populations to which it is administered, and studies have been carried out with the North American sample as a reference (Wenger et al., 2020). Adaptation and validation studies have been conducted for different populations (Berg et al., 2016).

Leiter International Performance Scale-Revised (Leiter-R)

This non-verbal instrument is widely used in children with speech/language, hearing, or motor impairments. The scale has been validated with the CTT and shown to have acceptable reliability. Studies to validate the instrument have been conducted with clinical populations. The instrument provides subtest, composite, percentile, and age equivalent scores.

The scale has been assessed in independent studies on the ASD population (Grondhuis & Mulick, 2013). These authors found a discrepancy in the scores with a control population and stated that the Leiter-R and the SB-5 may not be equivalent measures of intellectual functioning in these children.

Caudle et al. (2014) evaluated the concurrent validity of the Leiter-R and Vineland-II/WPPSI in hearing impaired children with cochlear transplants and found a positive correlation.

Vineland Adaptive Behavior Scales, Second Edition (Vineland-II)

Vineland-II is the gold standard for measuring adaptive behaviour. It has been evaluated with the CTT in a stratified North American sample of 3695 children. The scale comprises four domains which are in turn specified in 11 subdomains organised by items of increasing complexity. It offers raw and derived scores.

Some articles have compared the Vineland-II with tests for ASD or autism-specific scales. Yang et al. (2016), in agreement with the original Vineland research, found a distinct autism profile for scores equivalent to Vineland-II, but not in standard scores.

In terms of the instrument's concurrent validity, Scattone et al. (2011) compared standard scores on the Vineland-II and the Bayley-III. The authors concluded that the cognitive scores are statistically similar in both instruments, but that the standard composite scores for communication and motor skills were significantly higher in the Vineland-II.

Brazelton Neonatal Behavioural Assessment Scale (NBAS)

The NBAS was designed as a research instrument until it was later updated for clinical use. Administered in the first days of life, the scale is used both for the detection of deficits and for the identification of emerging abilities of newborn babies.

According to the search results, several studies use this scale as an instrument, but few measure its metric properties. As with the previous scales, studies are being conducted to administer the scale to different populations. However, as they are from previous years, they are fewer in number and of lower metric quality (Moragas et al., 2007; Costa et al., 2010; Başdaş et al., 2018). Lundqvist and Sabel (2000) determined if the scale detects behavioural differences in healthy newborns, as well as stress effects and individual characteristics such as gender. They found that girls showed higher levels of functioning than boys.

Neuropsychological Battery for Children, Second Edition (NEPSY-II)

The NEPSY-II is a battery of tests designed to assess the neuropsychological development of pre-schoolers and school age children. It has been evaluated with the CTT and shown good reliability and validity properties in a stratified sample of the 2003 US census. The NEPSY-II scores are divided into four categories: primary, process, contrast scores, and behavioural observations.

Several independent studies on the NEPSY-II have been conducted in young and adult populations, but metric reviews with children under 90 months of age are scarce. The only evidence can be found in Yao et al. (2018), who applied the Rasch IRT model to the Affect Recognition subtest and confirmed its appropriateness.

Revised Brunet-Lézine: Early Childhood Psychomotor Development Scale (BLR).

This scale assesses children's level of maturity in four domains and has been validated with the basic CTT parameters. The software provides the transformed scores, subjects' graphic profile, and a narrative report.

One article evaluated the concurrent validity between the Brunet-Lézine and Bayley-III scales for an older age group (18–24 months) (Cardoso et al., 2017). The Brunet-Lézine Scale is widely used in individuals with Down's syndrome, but no studies have measured its suitability for this population.

Discussion

One of the most limiting factors to improve the construction of developmental scales at the psychometric level is having access to representative population samples with homogeneous characteristics and well-differentiated groups (e.g., children with and without disabilities, types of disability, etc.). One way to increase the psychometric quality of the scales would be to increase the number of participants and group them according to homogeneous characteristics using non-incidental sampling.

In clinical practice, it is common to administer developmental scales to children with diverse types of developmental delays or disorders without the tests having the corresponding physical, temporal, and metric adaptations due to the lack of such adaptations. According to Silva et al. (2018), these scales mainly focus on disorders such as ASD and, to a lesser extent, cerebral palsy. Visser et al. (2012) highlighted the need to develop specific assessment instruments for different disabilities.

In this line, it is essential to validate the scales in the populations where they will be administered to avoid the cultural biases that occur when tests are translated without statistical validation. This is especially important in low-income countries, since most adaptation and validation processes are performed in North American and European samples. Also, in relation to the sample, it is essential to establish short age ranges taking into account the differential characteristics at each stage of development. It is generally

observed that ages are grouped together due to the need to adapt the sample to the mathematical assumptions of the analyses performed for validation.

In addition, conceptual concreteness and better sample quantity and quality would aid in the implementation of IRT. Most developmental scales are validated under CTT. These techniques have been superseded by new theories, particularly IRT. IRT analyses the properties of each item independently and provides information at different trait levels (Asún & Zúñiga, 2008). The invariance property facilitates the application of advanced psychometric techniques such as computerised tests, which are useful for selecting items according to the trait level of each subject (Muñiz, et al., 2005) and aid in early diagnosis.

This statistical procedure has been scarcely used to date in multidimensional developmental scales for children under 90 months of age. In this line, the MP-R Scale was the first international scale to be constructed and validated under IRT. Although this constitutes a significant step forward, statistical adjustments were made during the validation process. Thus, when IRT requires at least 150 subjects per population subgroup (López, 1995), it presents 150 participants "on average" in each of its subscales: the smaller groups have around 100 children and the larger ones almost 400, thus reducing the study validity. Furthermore, comparative analyses between subjects in different subgroups are made using CTT statistics, and tests are currently available for their analysis using IRT (Muñiz et al. 2005).

As for the scales analysed, few independent validation studies have been conducted for the MP-R. With respect to the Bayley-III, the MDI values are overestimated with respect to the previous version of the scale. Regarding the predictive validity of this scale, the results were found to vary by ages and by the tests/subtests, so moderate prognoses are recommended. The results of correlational studies generally focus on the cognitive, motor, and language domains. Battelle-II and PEDI show adequate metric robustness and independent IRT studies have been performed on these two scales. In addition, because PEDI is available in an extensively validated computerised version with an item bank, studies have included both children and adults with cerebral palsy. The studies on the Leiter-R focus on validation in populations

with communication difficulties. Although Vineland-II is considered one of the most notable tests at the international level (Silva et al, 2018), there is a marked scarcity of independent studies on this scale, especially compared to other scales of similar relevance. Vineland-II studies for subpopulations focus on children with ASD. The NEPSY-II scale has been studied in samples of older ages. The search results also showed that certain tests that were once fundamental for assessments in child diagnosis may be falling into disuse, such as the Brunet-Lézine Scale.

It is noteworthy that most of the scales have been assessed in cross-cultural studies with a view to adapting them to different nationalities and that all of them warn of the need to adapt the tests to the populations being assessed to ensure their validity, even in very young samples.

Limitations

The main limitations of this work are related to the use of different nomenclatures, which has made it difficult to search for the articles, the differences between countries in scale validation and use, and the scarcity of empirical evidence in many articles.

Future lines of research

To improve clinical practice, the tests must be shortened and their psychometric quality improved. To this end, statistical and technological advances must be applied, such as the current attempts to implement IRT in certain scales like the Merrill Palmer-R and validations in Nordic countries, or the effort to create software adapted to the PEDI, the PEDI-CAT. However, these developments are limited. In addition, examiners need to be trained on an on-going basis to ensure that they choose the most accurate scales for their purposes and are able to identify those that are best suited to their patients' conditions.

It is essential for researchers to continue to promote independent validation studies (AERA, APA, & NCME, 2014) of the scales to adapt them to different needs and populations and to certify their psychometric properties.

Conclusions

The overall results of this review show that multidimensional developmental scales are based on a broad conceptual framework with no apparent consensus. The methodological and metric validation of the scales is insufficient and could be improved. Moreover, few independent studies have assessed these scales. Psychometrics progress has been slow and heterogeneous (Silva et al., 2018) in adapting the methods and statistical theories used in common metrics to this specific field. The samples used to validate the scales tend to be small, and children under 12 months of age, with functional diversity, and from minority ethnic groups are under-represented. Further studies of these scales using larger and more homogeneous samples should be encouraged and the psychometric quality of their analysis and validation process should be improved.

Declaration of interests

The authors report no potential conflicts of interest.

References

Albuquerque, P., Guerra, M., Lima, M., Eickmann, S. (2018). Concurrent validity of the Alberta Infant Motor Scale to detect delayed gross motor development in preterm infants: A comparative study with the Bayley III. *Developmental Neurorehabilitation, 21*(6) 408-414. https://doi.org/10.1080/17518423.2017.1323974.

Alcantud- Marín, F. & Alonso- Esteban, Y. (2016). Predictive value of the Merrill-Palmer-R scale applied during the first year of life. *Educational psychology,22*(2), 87-92 doi: https://doi.org/10.1016/j.pse.2016.01.001

Adolph, K. E. & Hoch, J. E. (2019). Motor development: Embodied, embedded, enculturated, and enabling. *Annual Review of Psychology*, *70*(1), 141–164. https://doi.org/10.1146/annurev-psych-010418-102836

Amer, A., Kakooza-Mwesige, A., Jarl, G., Tumwine, J., Forssberg, H., Eliasson, A.-C., Hermansson, L. (2018). The Uganda version of the Pediatric Disability Assessment Inventory (PEDI-UG). Part II: Psychometric properties. *Child: Care, Health and Development*, *44*(4), 562-571. doi: https://doi.org/10.1111/cch.12562

Amundson, R., Kolobe, T., Arnold, S., McEwen, I. (2012). Concurrent validity of the school outcomes measures (SOM) and pediatric evaluation of disability inventory (PEDI) in preschool- aged children. *Phys Occup Ther Pediatr*, *35*(1), 40-53. doi: https://doi.org/10.3109/01942638.2014.975310

Anderson, P. J. & Burnett, A. (2016). Assessment of developmental delay in early childhood: concerns with the Bayley-III scales. *The clinical neuropsychologist,31*(2), 371-381. doi: https://doi.org/10.1080/13854046.2016.1216518

Asún, R., & Zúñiga, C. (2008). Advantages of the Politomic Models of Item Response Theory in the Measurement of Social Attitudes: The Analysis of a Case. *Psykhe, 17*(2), 103-115.

Başdaş, Ö., Erdem, E., Elmali, F., Kurtoğlu, S. (2018). The Brazelton neonatal behavioral assessment scale: A validity and reliability study in a Turkish sample. (2018). *Turkish Journal of Medical Sciences, 48*(2). doi: https://doi.org/10.3906/sag-1711-111.

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development (3rd ed.)*. San Antonio, TX: The Psychological Corporation.

Berg, M., Dolva, A., Kleiven, J., Krumlinde-Sundholm, L. (2016). Normative Scores for the Pediatric Evaluation of Disability Inventory in Norway. *Physical & Occupational Therapy in Pediatrics*. doi: https://doi.org/10.3109/01942638.2015.1050149

Bode, M., D'Eugenio, D., Mettelman, B., Gross, S. (2014). Predictive Validity of the Bayley, Third Edition at 2 Years for Intelligence Quotient at 4 Years in Preterm Infants. *Journal of Developmental & Behavioral Pediatrics, 35*(9), 570-5. doi: https://doi.org/10.1097/DBP.0000000000000110

Boonzaaijer, M., Suir I., Mollema, J., Nuysink, J., Volman, M., Jongmans, M. (2020). Factors associated with gross motor development from birth to independent walking: A systematic review of longitudinal research. *Child Care Health & Development, 47*(4), 525–561. doi: https://doi.org/10.1111/cch.12830

Brazelton, T., Nugent, J. (1997). *Scale for the evaluation of neonatal behavior*. Paidós.

Cardoso, F., Formiga, C., Bizinotto, T., Tessler, R., Blasbalg N., Francisco, R. (2017). Concurrent validity of the brunet-lézine scale with the bayley scale for assessment of the development of preterm infants up to two years. *Revista Paulista de Pediatría*, *35*(2), 144-150. doi: https://doi.org/10.1590/1984-0462/;2017;35;2;00005

Caudle, S., Katzenstein, J., S Oghalai, J., Lin, J., Caudle, D. (2014). Nonverbal cognitive development in children with cochlear implants:

relationship between the Mullen Scales of Early Learning and later performance on the Leiter International Performance Scales-Revised. *Assessment, 21*(1), 119–128. doi: https://doi.org/10.1177/1073191112437594

Committee on Children with Disabilities. (2001). Developmental surveillance and screening of infants and young children. *Pediatrics, 108*, 192–195. doi: http://pediatrics.aappublications.org/content/108/1/192

Costa, R., Figueiredo, B., Tendais, I., Conde, A., Pacheco, A., Teixeira, C. (2010). Brazelton Neonatal Behavioral Assessment Scale: A psychometric study in a Portuguese sample. *Infant Behavior & Development, 33*(4),510-7. doi: https://doi.org/10.1016/j.infbeh.2010.07.003

Dumas, H. M., Fragala-Pinkham, M. A., Rosen, E. L., & O'Brien, J. E. (2017). Construct validity of the pediatric evaluation of disability inventory computer adaptive test (PEDI-CAT) in children with medical complexity. *Disability and rehabilitation, 39*(23), 2446–2451. doi: https://doi.org/10.1080/09638288.2016.1226406

Elbaum, B., Gattamorta, K., Penfield, R. (2010). Battelle Developmental Inventory Assessment, 2nd Edition, Screening for use in state child outcomes measurement systems under the Individuals with Disabilities Education Act. *Early Intervention Magazine, 32*(4), 255–273. doi: https://doi.org/1053-4893; 1053-8151

Fernández, I., Álvarez, E., Estudi Llevant. (1989). *The psychomotor development of 1,702 children from 0 to 24 months.* University of Barcelona.

Floyd, R., Gathercoal, K., Roid, G. (2004). No evidence for ethnic and racial bias in the Tryout Edition of the Merrill-Palmer Scale-Revised. *Psychological Reports, 94* (1), 217-220. doi:https://doi.org/10.2466/pr0.94.1.217-220

Flynn, R., Huber, M., DeMauro, S. (2020). Predictive value of the BSID-II and the Bayley-III for cognitive function in early school age in very

preterm infants. *Global Pediatric Health*, *7*. doi: https://doi.org/10.1177/2333794x20973146

Gleason, M. M. (2010). Recognizing young children in need of mental health assessment: Development and preliminary validity of the Early Childhood Screening Assessment. *Infant Mental Health Journal*, *31*, 335–357. doi: https://doi.org/10.1002/imhj.20259

Goldin, R., Matson, J., Beighley, J., Jang, J. (2014). Autism spectrum disorder severity as a predictor of Battelle Developmental Inventory – Second Edition (BDI-2) scores in toddlers. *Developmental Neurorehabilitation*, *17*(1), 39–43. doi: https://doi.org/10.3109/17518423.2013.839585

Greene, M., Patra, K., Nelson, M., Silvestri, J. (2012). Evaluating preterm infants with the Bayley-III: Patterns and correlates of development. *Research in Developmental Disabilities, 33*(6), 1948–1956. doi: https://doi.org/10.1016/j.ridd.2012.05.024

Greene, M., Patra, K., Silvestri, J., Nelson, M. (2013). Re-evaluating preterm infants with the Bayley-III: Patterns and predictors of change. *Research in Developmental Disabilities, 34*(7), 2107–2117. doi: https://doi.org/10.1016/j.ridd.2013.04.001

Grondhuis, S. & Mulick, J. (2013). Comparison of the Leiter international performance scale - Revised and the Stanford-Binet Intelligence Scales, 5th Edition, in children with autism spectrum disorders. *American Journal on Intellectual and Developmental Disabilities, 118*(1), 44–54. doi: https://doi.org/10.1352/1944-7558-118.1.44

Haley, S., Coster, W., Ludlow, L., Haltiwanger, J., Andrellos, P. (1992). *Pediatric Evaluation of Disability Inventory (PEDI). Development, standardization and manual administration.* Trustees of Boston University.

Hanlon, C., Medhin, G., Worku, B., Tomlinson, M., Alem, A., Dewey, M., Prince, M. (2016). Adapting the Bayley Scales of infant and toddler development in Ethiopia: evaluation of reliability and validity. *Child:*

Care, Health and Development, 42(5), 699–708. doi: https://doi.org/10.1111/cch.12371

World Health Organization. (2012). *Developmental Difficulties in Early Childhood: Prevention, Early Identification, Assessment and Intervention in Low- and Middle-income Countries: A review*.

Jary, S., Whitelaw, A., Walløe, L., Thoresen, M. (2013). Comparison of Bayley-2 and Bayley-3 scores at 18 months in term infants following neonatal encephalopathy and therapeutic hypothermia. *Developmental Medicine & Child Neurology, 55*(11), 1053–1059. doi: https://doi.org/10.1111/dmcn.12208.

Johnson-Martin, N. M., Jens, Kenneth, G., Attermeier, S. M., & Hacker, B. J. (1994). Carolina Curriculum: Assessment and Exercises for Infants and Toddlers with Special Needs. Retrieved from: http://hdl.handle.net/11162/59068

Josse, D., Brunet, O., & Lézine, I. (1997). *Brunet Lèzine Revised: Early Childhood Psychomotor Development Scale.* Symtec.

Kamppi, D. & Gilmore, L. (2010). Assessing cognitive development in early childhood: A comparison of the Bayley-III and the Stanford-Binet- V. *The Australian Educational and Developmental Psychologist, 27*(02). doi: https://doi.org/10.1375/aedp.27.2.70

Karasik, L. B., & Robinson, S. R. (2022). Milestones or Millstones: How Standard Assessments Mask Cultural Variation and Misinform Policies Aimed at Early Childhood Development. Policy Insights from the Behavioral and Brain Sciences, *9*(1), 57–64. doi: https://doi.org/10.1177/23727322211068546

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY-II*. NCS Pearson Inc.

Krogh, M. & Væver, M. (2019a). Does gender affect Bayley-III scores and test-taking behavior? *Infant Behavior and Development.* doi: https://doi.org/10.1016/j.infbeh.2019.101352

Krogh, M. & Væver, M. (2019b). A longitudinal study of the predictive validity of the Bayley-III scales and subtests. *European Journal of*

Developmental Psychology, *16*(6), 727–738. doi: https://doi.org/10.1080/17405629.2018.1485563

Lalor, J. P., & Rodriguez, P. (2022). py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing.* doi: https://doi.org/10.1287/ijoc.2022.1250

Liao, P., & Campbell, S. (2004). Examination of the Item Structure of the Alberta Infant Motor Scale. *Pediatric Physical Therapy*, *16*(1), 31–38. doi: https://doi.org/10.1097/01.pep.0000114843.92102.98

Lin, L., Tu, Y., Yu, W., Ho, M., Wu, P. (2020). Investigation of fine motor performance in children younger than 36-month-old using PDMS-2 and Bayley-III. European. *Journal of Developmental Psychology*, *17*(5), 1–15. doi: https://doi.org/10.1080/17405629.2020.1732917

López, J. A. (1995). Estimation of parameters in TRI: A Bilog evaluation in small samples. *Psicothema, 7*(1), 173-185.

Lowe, J., Erickson, S., Schrader, R., Dunca, A. (2012). Comparison of the Bayley II Mental Developmental Index and the Bayley III cognitive scale: are we measuring the same thing? *Acta Paedratica,*101(2). doi: https://doi.org/10.1111/j.1651-2227.2011.02517

Lundqvist, C., Sabel, K. (2000). The Brazelton Neonatal Behavioral assessment scale detects differences among newborn infants of optimal health. *Journal of Pediatric Psychology*, *25*(8), 577-582. doi: https://doi.org/10.1093/jpepsy/25.8.577

Månsson, J., Stjernqvist, K., Serenius, F., Ådén, U., Källén, K. (2019). Agreement Between Bayley-III Measurements and WISC-IV Measurements in Typically Developing Children. *Journal of Psychoeducational Assessment*, *37*(5), 603-616. doi: https://doi.org/10.1177/0734282918781431

Månsson, J., Källén, K., Eklöf, E., Serenius, F., Ådén, U., Stjernqvist, K. (2021). The ability of Bayley-III scores to predict later intelligence in children born extremely preterm. *Acta Paediatrica missing data in this reference.* doi: https://doi.org/10.1111/apa.16037

Mayrand, L., Mazer, B., Menard, S., Chilingaryan, G. (2009). Screening for motor deficits using the pediatric evaluation of disability inventory (PEDI) in children with language impairment. *Developmental Neurorehabilitation*, *12*(3), 139–145. doi: https://doi.org/10.1080/17518420902936722

McCoy, D. C. (2022). Building a model of cultural universality with specificity for global early childhood development. *Child Development Perspectives*, *16*, 27–33. doi: https://doi.org/10.1111/cdep.12438

Milne, S., Mcdonald, J. Comino, E. (2015). Alternate scoring of the Bayley-III improves prediction of performance on Griffiths Mental Development Scales before school entry in preschoolers with developmental concerns. *Child: Care, Health and Development*, *41*(2), 203–212. doi: https://doi.org/10.1186/s12887-015-0457-x

Moore, T., Johnson, S., Haider, S., Hennessy, E., Marlow, N. (2012). Relationship between test scores using the second and third editions of the bayley scales in extremely preterm children. *The Journal of Pediatrics*, *160* (4), 553–558. doi: https://doi.org/10.1016/j.jpeds.2011.09.047

Moragas, C., Deu, A., Mussons, F., Costa, E., Zurita, M. (2007). Psychometric evaluation of the Brazelton Scale in a sample of Spanish newborns. *Psicotema*, *19*(1), 140-9.

Morsan, V., Fantoni, C., Tallandini, M. (2018). Age correction in cognitive, linguistic, and motor domains for infants born preterm: an analysis of the Bayley Scales of Infant and Toddler Development, Third Edition developmental patterns. *Developmental Medicine & Child Neurology*, *60*(8), 820-825. doi: https://doi.org/10.1111/dmcn.13735

Muñiz, J., Fidalgo, A., García-Cueto, E., Martínez, R., Moreno, R. (2005*). Analysis of the items.* The Wall.

Muñiz, J. (2010). Test of Theories of the Classical Theory and Theory of Responses to Items. *Papeles del Psicólogo*, *31*(1), 57-66.

Newborg, J. (2005). *Battelle Developmental Inventory - second edition.* Riverside.

Nishijima, M., Yoshida, T., Matsumura, K., Inomata, S., Nagaoka, M., Tamura, K., Makimoto, M. (2021). Correlation between the Bayley-III at 3 years and WISC-IV at 6 years. *Pediatrics International missing data in this reference*. doi: https://doi.org/10.1111/ped.14872

Nitsana, M. (2011). *The Battelle Developmental Inventory, 2nd Edition: A study of Concurrent Validity and Stability in Young Children with Known Disabilities Chairperson missing data in this reference*, Umi Dissertation Publishing.

Nordmark, E., Jarnlo, G., Hägglund, G. (2000). Comparison of the Gross Motor Function Measure and Paediatric Evaluation of Disability Inventory in assessing motor function in children undergoing selective dorsal rhizotomy. *Developmental Medicine & Child Neurology*, *42*(4), 245–252. doi: https://doi.org/10.1017/s0012162200000426

Olusanya, B., Hadders-Algra, M., Breinbauer, C., Williams, A., Newton, C. R., & Davis, A. (2021). The conundrum of a global tool for early childhood development to monitor SDG indicator 4.2.1. The Lancet Global Health, *9*(5), e586– e587. doi: https://doi.org/10.1016/S2214-109X(21)00030-9

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Moher, D. (2021). *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews.* International Journal of Surgery *10*, 88. doi: https://doi.org/10.1186/s13643-021-01626-4

Peters, M. (2013). *Determining the Clinical Utility of the Merrill-Palmer-Revised Scales of Development in a Sample of Children with Autistic Disorder*. George Fox University.

Peyton, M., Wroblewski, K., Rogers, E., Kohn, M., Glass, H (2020). Simultaneous validity of the Warner Assessment of Early Development of Functional and Adaptive Skills and the Bayley Scales

of Infant and Toddler Development, Third Edition. *Developmental medicine and child neurology,63*(3), 349–354. doi: https://doi.org/10.1111/dmcn.14737

Roid, G. & Miller, L. (1997). *LEITER-R International Manipulative Scale*. Psymtec.

Roid, G., Sampers, J., Anderson, G., Erickson, J., Post, P. (20004). *Merrill-Palmer- Revised. Scales of Development.* Stoelting Company.

Scattone, D., Raggio, D., May, W. (2011). Comparison of The Vineland adaptive behavior scales, second edition, and the Bayley scales of infant and toddler development, third edition. *Psychological Reports, 109*(2), 626–634. doi: https://doi.org/10.2466/03.10.PR0.109.5.626-634

Sharp, M., & DeMauro, S. (2017). Counterbalanced comparison of the BSID-II and Bayley-III at eighteen to twenty-two months corrected age. *Journal of Developmental & Behavioral Pediatrics, 38* (5), 322–329. doi: https://doi.org/10.1097/DBP.0000000000000441

Silva, M., Mendonça, F., Euclides J., Mônego, B., Bandeira, D. (2018). Instruments for multidimensional assessment of child development: a systematic review. *Early Child Development and Care*, 1–15. doi: https://doi.org/10.1080/03004430.2018.1528243

Sipes, M., Matson, J., Turygin, N. (2011) The use of the Battelle Developmental Inventory - Second Edition (BDI-2) as an early evaluator of autism spectrum disorders. *Developmental neurorehabilitation, 14*(5), 310–314. doi: https://doi.org/10.3109/17518423.2011.598477

Sparrow, S., Cicchetti, D., & Balla, D. (2005). *Vineland Adaptive Behavior Scales (2nd ed.).* Circle Pines.

Spittle, A., Spencer-Smith, M., Eeles, A., Lee, K., Lorefice, L., Anderson, P., Doyle, L. (2013). Does the Bayley-III Motor Scale at 2 years predict motor outcome at 4 years in very preterm children? *Developmental*

Medicine & Child Neurology, *55*(5), 448–452. doi: https://doi.org/10.1111/dmcn.12049

Torras-Mañá, M., Guillamón-Valenzuela, M., Ramírez-Mallafré, A., Brun-Gasca, C., Fornieles-Deu, A. (2014). Usefulness of the Bayley scales of infant and toddler development, third edition, in the early diagnosis of language disorder. *Psicotema*, *26*(3), 349-56. doi: https://doi.org/10.7334/psicothema2014.29

Torras-Mañá, M., Gómez-Morales, A., González-Gimeno, I., Fornieles-Deu, A., Brun-Gasca, C. (2016). Assessment of cognition and language in the early diagnosis of autism spectrum disorder: Usefulness of the Bayley Scales of infant and toddler development, third edition. *Journal of Intellectual Disability Research*, *60*(5), 502–511. doi: https://doi.org/10.1111/jir.12291

Vierhaus, M., Lohaus, A., Kolling, T., Teubert, M., Keller, H., Fassbender, I., Freitag, C., Goertz, C., Graf, F., Lamm, B., Spangler, S. M., Knopf, M., Schwarzer, G. (2011). The development of 3- to 9-month-old infants in two cultural contexts: Bayley longitudinal results for Cameroonian and German infants. European *Journal of Developmental Psychology*, *8*(3), 349–366. doi: https://doi.org/10.1080/17405629.2010.505392

Visser, L., Ruiter, S., van der Meulen, B., Ruijssenaars, W., Timmerman, M. (2012). A Review of Standardized Developmental Assessment Instruments for Young Children and Their Applicability for Children with Special Needs. J*ournal of Cognitive Education and Psychology*, *11*(2), 102–127. doi: https://doi.org/10.1891/1945-8959.11.2.102

Vos-Vromans, D., Ketelaar, M., Gorter, J. (2005). Responsiveness of evaluative measures for children with cerebral palsy: The Gross Motor Function Measure and the Pediatric Evaluation of Disability Inventory. *Disability & Rehabilitation, 27*(20), 1245–1252. doi: https://doi.org/10.1080/09638280500076178

Wenger, I., Schulze, C., Kottorp, A. (2020). Are the American normative standard scores applicable to the German version of the Pediatric

Evaluation of Disability Inventory (PEDI-G)? *Scandinavian Journal of Occupational Therapy*, *28*(2), 1–11. doi: https://doi.org/10.1080/11038128.2020.1726452

Yang, S., Paynter, J., Gilmore, L. (2016). Vineland Adaptive Behavior Scales: II Profile of Young Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *46*(1), 64–73. doi: https://doi.org/10.1007/s10803-015-2543-1

Yao, S., Bull, R., Khng, K., Rahim, A. (2018). Psychometric properties of the NEPSY-II affect recognition subtest in a preschool sample: a Rasch modeling approach. *The Clinical Neuropsychologist*, 1-18. doi: https://doi.org/10.1080/13854046.2017.1343865

Sara M. Luque de Dios University of Córdoba, Spain
ORCID: 0000-0002-8931-406X
Araceli Sánchez-Raya University of Córdoba, Spain
ORCID: 0000-0001-6264-3456
Juan A. Moriana University of Córdoba, Spain
ORCID: 0000-0003-0577-821X

Contact Address: ed1saram@uco.es