Psychological Assessment

Psicologia: Teoria e Prática, 26(2), ePTPPA14444. São Paulo, SP, 2024.
ISSN 1516-3687 (print), ISSN 1980-6906 (electronic version). **Evaluation system:** double-blind review.
https://doi.org/10.5935/1980-6906/ePTPPA14444.en
Universidade Presbiteriana Mackenzie

Reliability of the BAYLEY Scale: Assessing Children Affected by Zika Virus

Tamiles Cerqueira Lopes da Silva[1], George Anderson Alves dos Santos[1], Leticia Marques dos Santos[2], Vanessa Madaschi[3], Andrea Perosa Saigh Jurdi[4], Darci Neves dos Santos[1]

[1] Federal University of Bahia (Universidade Federal da Bahia [UFBA]), Salvador, BA, Brazil
[2] Federal University of Bahia (Universidade Federal da Bahia [UFBA]), Salvador, BA, Brazil
[3] Mackenzie Presbyterian University (Universidade Presbiteriana Mackenzie [UPM]), São Paulo, Brazil
[4] Federal University of São Paulo (Universidade Federal de São Paulo [Unifesp]), São Paulo, SP, Brazil

Received: April 2nd, 2021.
Accepted: September 20th, 2023.
Section editor: Ana Alexandra Caldas Osório.

Author Note

Tamiles Cerqueira Lopes da Silva https://orcid.org/0000-0002-4769-640X
George Anderson Alves dos Santos https://orcid.org/0000-0002-5600-5287
Leticia Marques dos Santos Silva https://orcid.org/0000-0001-5963-2166
Vanessa Madaschi https://orcid.org/0000-0002-6954-4407
Andrea Perosa Saigh Jurdi Silva http://orcid.org/0000-0002-1111-5562
Darci Neves dos Santos https://orcid.org/0000-0003-1500-8294

Correspondence concerning this article should be addressed to Tamiles Cerqueira Lopes da Silva, Rua Basilio da Gama, Canela, Salvador, BA, Brazil. CEP 40110040. Email: tamiles.live@hotmail.com

Abstract

Understanding and measuring child development's determinants can contribute to better healthcare guidance during early childhood. The Bayley-III Scales were used to assess the development of children with and without a diagnosis of Zika Virus Congenital Syndrome participating in a longitudinal study conducted in their homes. Inter-observer reliability was examined at three assessment points, and the training and supervision process of the interdisciplinary team was also described to standardize the instrument's application procedures in data collection. Reliability measures were produced using the Kappa index and the Intraclass Correlation Coefficient (ICC). The average Kappa values were 0.92, 0.89, and 0.96 for the first, second, and third reliability measurements between assessors. The ICC remained above 90% for each subscale assessed in the three measurements. The results show excellent reliability indicators when applying the scales, suggesting the importance of team training and supervision to ensure an inter-observer reliability standard in assessing children with neurodevelopmental disorders in population studies.

Keywords: Reliability Measures, neurodevelopmental disorders, Bayley-III Scales, development measures, test accuracy, neuropsychological assessment, child development

CONFIABILIDADE NA APLICAÇÃO DA ESCALA BAYLEY: AVALIANDO CRIANÇAS ACOMETIDAS PELO ZIKA VÍRUS

Resumo

Entender e mensurar determinantes que influenciam o desenvolvimento infantil, poderá contribuir para um melhor direcionamento dos cuidados de saúde na primeira infância. Utilizaram-se as escalas Bayley III para avaliar o desenvolvimento de crianças com e sem diagnóstico da síndrome congênita do zika vírus, participantes de um estudo longitudinal avaliados em domicílio. Examinou-se a confiabilidade interobservadores em três pontos de avaliação, descrevendo também o processo de treinamento e supervisão da equipe interdisciplinar, para padronização dos procedimentos de aplicação do instrumento na coleta de dados. Produziram-se medidas de confiabilidade pelo índice kappa e coeficiente de correlação intraclasse (CCI). Os valores médios de kappa corresponderam a 0,92, 0,89 e 0,96, respectivamente, para a primeira, segunda e terceira medidas de confiabilidade entre aplicadores. O CCI se manteve acima de 90% para cada subescala avaliada nas três medidas realizadas. Os resultados demonstram excelentes indicadores de confiabilidade na aplicação das escalas, sugerindo a importância do treinamento e supervisão da equipe para conferir um padrão de confiabilidade interobservadores da avaliação de crianças com transtornos do neurodesenvolvimento em estudos populacionais.

Palavras-chave: medidas de confiabilidade, transtornos do neurodesenvolvimento, escalas Bayley III, precisão do teste, avaliação neuropsicológica, desenvolvimento infantil.

FIABILIDAD DE LA ESCALA BAYLEY: EVALUACIÓN DE LOS NIÑOS AFECTADOS POR EL VIRUS ZIKA

Resumen

Comprender y medir los factores determinantes que influyen en el desarrollo infantil puede contribuir a una mejor orientación de la atención médica en la primera infancia. En este estudio longitudinal, se utilizaron las Escalas Bayley III para evaluar el desarrollo de niños con y sin diagnóstico de la Síndrome Congénita del Virus Zika, quienes fueron evaluados en sus hogares. Se examinó la fiabilidad entre observadores en tres puntos de evaluación, describiendo también el proceso de capacitación y supervisión del equipo interdisciplinario para estandarizar los procedimientos de aplicación del instrumento en la recopilación de datos. Se calcularon las medidas de fiabilidad mediante el índice Kappa y el Coeficiente de Correlación Intraclase (CCI). Los valores promedio de Kappa fueron 0,92, 0,89 y 0,96 para la primera, segunda y tercera medida de fiabilidad entre los aplicadores. El CCI se mantuvo por encima del 90% para cada subescala evaluada en las tres medidas realizadas. Los resultados muestran excelentes indicadores de fiabilidad en la aplicación de las escalas, lo que sugiere la importancia de la capacitación y supervisión del equipo para garantizar un estándar de fiabilidad entre observadores en la evaluación de niños con trastornos del neurodesarrollo en estudios poblacionales.

Palabras-clave: Medidas de fiabilidad, trastornos del neurodesarrollo, Escalas Bayley III, precisión del test, evaluación neuropsicológica, desarrollo infantil.

It is well-established that several factors can influence child development. Understanding and systematically quantifying these factors can contribute to more effectively targeting healthcare efforts. The American Academy of Pediatrics (Sandler et al., 2001) recommends assessing and monitoring at-risk infants and young children throughout early childhood, allowing for early and specific interventions in potential neurodevelopmental alterations. Population studies that assess children with developmental disorders challenge the scoring patterns used to measure skill acquisition due to the quantitative nature of the instruments. These instruments measure skill acquisition across age groups and utilize normative samples from populations with typical development (Rodrigues, 2012).

Developmental delays can have various causes. It is worth noting that Congenital Zika Virus Syndrome (CZVS) encompasses a range of congenital anomalies that can include visual, auditory, and neuropsychomotor alterations in individuals exposed to this infection during pregnancy (Ministry of Health, 2016). The severity of these alterations can vary, posing challenges for healthcare professionals involved in monitoring and assessing child development.

The first step in ensuring attention and access to specific programs is through diagnosis, with early screening, identification, and appropriate referral being essential. Strong evidence underscores the importance of early intervention (Zwaigenbaum & Penner, 2018), emphasizing the use of reliable measures with high levels of sensitivity, specificity, and reliability (Campos et al., 2006; Santos & Ravanini, 2006; Blair & Hall, 2006).

For children with complex developmental issues, procedures ranging from surveillance to screening for risk factors, as well as assessments to determine functional diagnoses, are essential pillars for ensuring compliance with the recommendations of the Ministry of Health recommendations and early intervention when necessary. The clinical reasoning process is initiated based on the assessment to determine the best intervention plan (Gourladin & Sá, 2022).

However, the availability of standardized and culturally validated scales in our language poses a challenge in assessing children with developmental disorders (Visser et al., 2014; Madaschi et al., 2016). Assessing individuals with disabilities that impact all domains of child development requires test accommodations to enable the individual's full participation in the process (Bayley, 2006). Accommodations refer to changes in the standard test administration procedures to overcome the functional limitations of the participant, thereby increasing the validity of inferences drawn from the scores obtained. It is considered relevant to address the functional limitations individuals may experience when attempting to demonstrate proficiency in an assessment (Kettler, 2012).

Initially published in 1969, the Bayley Scales, in their original American version, are considered the gold standard for meeting all psychometric properties (Diamond, 2000). After nearly 50 years of research, version IV of the instrument is currently available, maintaining its excellent quality and meeting rigorous psychometric properties. The Bayley-III Scales (BSID III) used in this research is the most recent version available in Brazil (Madaschi & Paula, 2011). It is

also one of the instruments recommended by the Early Stimulation guidelines resulting from Microcephaly (Ministry of Health, 2016) to identify developmental delays, plan interventions, and document progress and evolution (Bayley, 2006). Although standardized, it offers flexibility in application by considering the inherent dynamism of various infant assessment situations (Bayley, 2006), making it suitable for assessing cognitive, linguistic, and motor skills in children affected by congenital anomalies.

This version of the scale was adopted as an assessment instrument in the Research Project "Effects of Congenital Neurological Manifestations Associated with the Zika Virus on Child Development: A Prospective Cohort Study in the Context of Primary Care in Salvador-BA" to assess developmental consequences in children born during the epidemic, with a focus on home-based assessment (Santos et al., 2022).

Given the potential effects of interviewers on the reliability of the obtained responses, the research design included procedures to measure agreement among assessors throughout the study. Confidence in the results is partially a function of the amount of disagreement or error introduced into the study due to inconsistencies among instrument administrators. Reliability is dynamic and depends on the instrument's function, the population in which it is administered, circumstances, and context. These factors highlight the importance of ongoing training and supervision to achieve adequate inter-rater reliability (Souza et al., 2017). Equivalence reliability allows for identifying the extent to which assessors could observe and measure the phenomenon or variable appropriately and as predicted by the instrument's validity.

Measuring agreement among collectors refers to stability, internal consistency, and measure equivalence, although reliability is not a fixed property of the instrument (Souza et al., 2017).

Therefore, it is imperative to subject an assessment team to a training and test-retest process to measure the level of equivalence among them, aiming to minimize measurement errors. Equivalence refers to the degree of agreement between two or more observers regarding the scores of an instrument (Heale & Twycross, 2015). Internal consistency among administrators is expressed by the Kappa coefficient (K), which measures the degree of agreement between proportions derived from dependent samples (Cohen, 1968). The Intraclass Correlation Coefficient (ICC), a measure assessing the reliability or consistency among multiple measurements made by different administrators, is included.

Given the significance of developmental assessment for the therapeutic intervention process with children and their families, this article examines the reliability obtained at three assessment points in a longitudinal study using the Bayley-III Infant Development Scales. Furthermore, it describes the continuous training and supervision of the interdisciplinary team, aiming to standardize the application procedures in data collection.

Method

A longitudinal quantitative, observational, and descriptive study to assess agreement among administrators of the Bayley-III Infant Development Scales in Salvador (BA), Brazil's fourth most populous capital city.

Participants

The reliability design model chosen for this study consisted of balanced incomplete blocks, as Fleiss (1981) described, in which one examiner interviews while the other observes the examination as a neutral spectator. According to the author, Balanced incomplete blocks refer to a specific type of experimental design used in studies involving the assessment of multiple treatments. In this design, each participant or experimental unit is not exposed to all possible combinations of treatments but rather to a subset of them. This approach is beneficial when dealing with a high total number of combinations, making it impractical to test all of them.

The method of balanced incomplete blocks allows for reducing the size of the experiment and conserving resources while maintaining the balance between the tested conditions. To ensure the validity of the results, the allocation of treatments to participant blocks must be random or systematic, depending on the adopted strategy. This controls external variability and improves the precision of conclusions, making the study more robust and reliable.

After theoretical training, pairs consisting of two administrators recorded scores based on the same interview but conducted independent assessments. The simple arrangement method, using combinatorial analysis, was employed to form these pairs. At another time, the roles of the pairs were reversed when assessing another child. Reliability measures for the three assessment points of the longitudinal study were obtained independently. The level of agreement between assessors was measured for each assessment point of the longitudinal study, with three independent measures of cognitive, motor, and linguistic performance in children with and without exposure to CZVS.

Team Composition

A team was formed with students from the Interdisciplinary Bachelor Programs in Health, Psychology, Public Health, and Physiotherapy, as well as professional psychologists. The first measurement occurred at the baseline between April 2017 and March 2018, with six interviewers conducting 29 instrument applications following the above-described design to assess reliability. A second measurement was conducted between May 2018 and March 2019 when eight new interviewers joined one existing member, and these nine members conducted 67 applications for the reliability study. The final measurement occurred between February and August 2019, during which four administrators conducted 11 assessments for the reliability sample.

Instrument

The Bayley-III Scales aim to measure the performance of infants and young children's performance, identify competencies and critical points, and contribute to proper therapeutic intervention planning. Five domains are investigated through direct child assessments, addressing cognitive, expressive, and receptive language, gross and fine motor skills, and socioemotional and adaptive behavior scales applied in interviews with parents (Bayley, 2006).

According to the instrument's Technical Manual, the administration time can vary from 50 minutes for children up to 12 months to 90 minutes for those over 13 months.

Its structure provides five types of scores: (i) raw total scores, (ii) scaled scores, (iii) composite scores, (iv) percentile-based rankings, and (v) developmental scores. For each domain, the raw score is defined as the total number of items for which the child receives credit, summed with the number of items before the child's starting point. The scaled score is derived from the raw score. It ranges from 1-19, with an average of 10 and a standard deviation of 3, while the composite score is calculated based on the scaled score and ranges from 40-160, with an average of 100 and a standard deviation of 15.

Among the effects caused by CZVS, inadequate development of gross motor skills stands out, affecting the child's ability to roll, sit, and, in many cases, maintain control of their head. Regarding fine motor skills development, difficulties in performing manual activities are reported. In the sensory system, compromised visual and auditory capabilities are observed, resulting in severe difficulty understanding and producing language (Wheeler, 2018). However, the BSID III can be adjusted in its standard version, which allows for streamlining assessment-related tasks, provided there are no changes in their content and objectives (Visser et al., 2013; Visser et al., 2014).

Selection of the data collection team

At the beginning of the project, a call for applications was launched to offer an introductory course on the Bayley-III Infant Development Scale, followed by the selection of assessors based on their demonstrated performance and level of engagement with the project. Considering the assessment points of the longitudinal study, the age limit of 42 months for including the child, and the fluctuation of interviewers throughout the data collection, new calls were opened, transforming the training offering into a certified extension course by the Universidade Federal da Bahia (UFBA, Federal University of Bahia).

Training of Assessors

In its three stages, the training content covered topics related to typical and atypical child development, as well as the understanding and use of the Bayley-III Scales.

Two psychology professors with experience in child development, epidemiological aspects of development, and quantitative assessment instruments conducted the training and supervision process. This process also included a third psychologist responsible for coordinating

the team's activities during home visits or at health facilities for data collection throughout the study period.

The first practical training activity involved assessing children known to the team, conducted in their own homes, followed by a pilot experience involving children with typical and atypical development. Once the theoretical and practical requirements were met, the assessors started their activities with study participants in the research setting.

Initially, all team members were paired with a more experienced partner in a double role: one member approached the child and recorded the scores, while the second member observed the approach and independently and silently recorded the scores. Subsequently, these assessments were discussed in weekly supervision meetings to assess the performance of team members who were being trained. As the assessments progressed during the second and third measurements in the cohort, regular supervision meetings gave way to scheduled meetings based on the assessors' needs.

Data Collection

The workflow began with the team leaving the Institute of Collective Health, reaching the assessment location (home or, exceptionally, a health facility due to safety concerns regarding the team), administering the instrument, and returning to the Institute. The duration of each assessment with the instrument was approximately two hours, varying according to the child's developmental profile, health and well-being fluctuations, and the need for occasional breaks based on individual characteristics. The team observed that home-based assessment made the process more comfortable for the child, allowing the assessor to become familiar with existing limitations, improving communication, and providing the necessary flexibility to implement the required test accommodations.

From the initial assessments of children with confirmed CZVS diagnoses, difficulties in administering the Bayley-III Scales were observed, potentially disadvantaging the child due to the administration of the test per se. It was recognized that accommodations would be necessary alterations to standard procedures to overcome individuals' functional impairments and increase response validity (Kettler, 2012). It was also noted that the manual (Bayley, 2006) established criteria and necessary rigor for adaptations. Despite being scarce, studies that applied the BSID III with facilitations to assess children with multiple impairments show that the use of facilitations corrected differences in raw test scores, especially in the cognitive scale, increasing the validity and use of this instrument under these conditions (Ruiter et al., 2010). Based on this evidence, a tool was developed to assess the subject's cognitive abilities with the best possible expression and precision of their development. The scale was standardized through training and supervision during the study.

Data analysis

An equivalence analysis was conducted to identify the degree of agreement between pairs of observers concerning the scores on the Bayley Scale (Heale & Twycross, 2015). The intraclass correlation coefficient was employed to calculate this agreement when the results were continuous variables. This coefficient is widely used to assess agreement between repeated measures, especially when multiple observers are involved. Results are considered excellent when the agreement between values exceeds 0.75.

The Kappa index (K) was chosen to analyze the agreement between categorical variables to measure the degree of agreement between proportions derived from dependent samples (Cohen, 1968). The obtained values are classified as follows: ≤ 0, no agreement; 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, excellent agreement (McHugh, 2012; Souza et al., 2017). Statistical analysis was performed using the Statistical Package for the Social Sciences (SPSS version 20) and *R (R Statistical Language version 3.6.1).

Results

The first reliability assessment measurement conducted at the baseline involved six assessors and 29 children assessed by pairs of assessors, yielding the following results. Kappa values ranged from 0.47 to 1, with an average of 0.92. At least 70% of the questions showed values close to 1, indicating excellent agreement between assessors. The average of the 58 scores from the 29 assessments at this point was 84.2 (SD = 24.8) for the cognitive scale, 81.3 (SD = 25.4) for language, and 80.7 (SD = 28.8) for motor skills.

The second measurement involved 67 children and nine assessors for equivalence testing. Kappa values ranged from 0.43 to 1, with an average of 0.89. Approximately 70% of the questions showed Kappa values close to 1, maintaining excellent agreement between assessors. For the 67 assessments at this point, the average of the 134 scores obtained was 73.3 (SD = 22.7) for the cognitive scale, 72.2 (SD = 24.7) for the language scale, and 70.0 (SD = 27.5) for the motor scale.

In the third measurement, 11 children comprised the equivalence sample with the participation of four assessors. Kappa values ranged from 0.21 to 1, with an average of 0.96. Seventy percent of the questions had Kappa values close to 1, demonstrating excellent agreement between assessors. In this final measurement, 11 assessments (22 scores) were conducted with averages of 63.9 (SD = 17), 57.9 (SD = 19), and 58.1 (SD = 26) for the cognitive, language, and motor scales, respectively. Table 1 displays the distribution of Kappa coefficient ratings for the three measurements in the study.

The analysis of reliability measures for applying the Bayley-III Child Development Scale in this longitudinal study demonstrated satisfactory Kappa coefficients, indicating excellent agreement between assessors, with a slight increase in average values in the first and third measurements (Table 1).

Notably, the items with low agreement were distributed as follows in this longitudinal reliability study. At the baseline, only one item in the cognitive domain (Item 8) showed low agreement. The second measurement recorded six items with low agreement between assessors, including two in Expressive Language (1;43), two in Receptive Language (2;6), one cognitive item (4), and one Fine Motor item (1). In the third measurement, only two items related to Expressive Language (2;6) showed low agreement.

Table 1

Concordance obtained via Kappa Coefficient for three reliability assessment measurements among assessors in the longitudinal study, April 2017 – August 2019, Salvador (BA), Brazil

Concordance	No. of questions	%	Kappa value
First assessment: baseline			
Moderate	1	0.31	0.41 to 0.60
Substantial	36	11.04	0.61 to 0.80
Excellent	289	88.65	0.81 to 1.00
Second assessment			
Moderate	10	3.37	0.41 to 0.60
Substantial	47	15.82	0.61 to 0.80
Excellent	238	80.13	0.81 to 1.00
Third assessment			
Considerable	2	0.70	0.21 to 0.40
Substantial	24	8.42	0.61 to 0.80
Excellent	259	90.88	0.81 to 1.00

Source: own authorship
Note: Concordance categories without frequency were omitted.

The intraclass correlation coefficient for each of the three performance scores at each assessment point remained above 90%, indicating a high level of agreement between assessors, also classified as excellent (Table 2).

Table 2

Intraclass Correlation Coefficients (ICC) for reliability assessments according to cognitive, language, and motor subscales conducted at three assessment points in the longitudinal study, April 2017 – August 2019, Salvador (BA), Brazil

Assessments According to Scales		*CCI*	*CCI* 95%
First assessment: baseline			
	Cognitive	0.925	0.865
	Language	0.951	0.907
	Motor	0.939	0.889
Second assessment			
	Cognitive	0.963	0.942
	Language	0.994	0.991
	Motor	0.980	0.968
Third assessment			
	Cognitive	0.998	0.995
	Language	0.997	0.992
	Motor	1.000	0.999

Source: own authorship

Discussion

In a longitudinal study with a turnover of assessor teams, it was possible to maintain data quality, reproducing consistent results over time and space, as demonstrated by the reliability measures obtained. The initial training format was supplemented by regular supervision, which qualified the team by identifying doubts and disagreements, thus establishing a high inter-rater reliability standard in the three assessment measurements (Souza et al., 2017).

Analyzing the overall results of the three assessed points, some variations in reliability levels were observed, indicating the importance of conducting reliability assessments between assessors for each new assessment point, even if the instrument in question had previously demonstrated high reliability.

The increasing number of recent national studies that have used the Bayley-III Scales indicates the importance and utility of this instrument in diagnosing motor, cognitive, and language delays in Brazilian children (Ferreira et al., 2014; Hentges et al., 2014). Investigations examining findings for populations with complex developmental disorders are scientifically relevant, emphasizing the need to test the necessary accommodations to maintain evidence of validity and reliability of specific assessment tools. This study faced an essential challenge in assessing children at high risk of developmental delays due to multiple disabilities and frequent global developmental impairment.

Accommodations are required in such situations, allowing for modifications to the standard test administration procedures to overcome the participant's functional deficiencies, thus increasing the validity of inferences based on obtained scores. The functional impairment the subject may experience when attempting to demonstrate proficiency in an assessment is considered relevant (Kettler, 2012). An effort was made to enable psychological evaluation in a population with many limitations, which could otherwise render the subjects untestable by other instruments. This contributes to advancements in child development assessment in Brazil, without any similarly validated tool (Madaschi et al., 2016).

The BSID-III was administered using standard procedures, with adaptations for the child's visual or motor impairments, as suggested by the manual (Bayley, 2006). Among the accommodations used in our study, we refer to some examples, such as using light and brightness, increasing the size of manipulative objects, and extending the time for the child to respond.

Additional test accommodations in the application of the BSID-III have been used. For example, ceiling lights were turned off for children with clear vision, and a flashlight was used to provide contrast (Wheeler et al., 2020). Although the item scope covers the developmental spectrum from birth to early childhood, the dependence on visual and motor production can penalize children with CZVS in demonstrating what they can do. On the other hand, raw BSID-III and age-equivalent scores provide a sensitive measure of potential change over time, allowing for monitoring skill gains or losses in response to time, treatment, or seizures (Wheeler et al., 2020).

While at least 70% of the questions analyzed had Kappa values close to 1, questions were distributed across the three assessments with low agreement among assessors. The baseline assessment occurred with questions from the cognitive domain and the other points, as well as language and motor domains. One possible hypothesis explaining the low agreement is the potential cognitive domain impairment due to viral infection and the high correlation between these three domains. Knowing that motor impairment due to axial and appendicular hypertonia can affect the magnitude of the child's response or behavior, it could make it difficult for the assessor to judge the application of these tests that deviated from high agreement. In this regard, greater attention is recommended in training for items that require greater assessor sensitivity to recognize the child's performance.

Regarding items of expressive language with low agreement, tests involving undifferentiated guttural sounds require familiarity for the assessor to interpret and score. Finally, low agreement in this study reached items of subjective dimensions, such as the social smile response when talking to the child. Despite the various potential sources of disagreement in applying a psychometric test, the scores obtained in this study were highly consistent, reflecting uniformity in approaching children and interpreting their responses. It is considered possible to deal with factors related to the instrument, population, and context, achieving a high level of reliability (Souza et al., 2017). Using the equivalence type's reliability, the assessors' aptitude to observe and measure the phenomenon appropriately, as recommended by the Bayley-III Manual for Child Development Scale, was identified.

Assessing children in their homes or occasionally at their reference Health Unit facilitated their approach. The assessor's access to this space favored some familiarity with the child's limitations, improved communication, and the necessary flexibility to adopt accommodations. Ecological validity emphasizes a new understanding of the relationship between assessment results and the performance of daily tasks. It also considers the development of tests composed of everyday cognitive functions so that inferences can be quickly drawn from the results and the individual's probable ability to perform those tasks in daily life (Spooner & Pachana, 2006). According to Pasquali (2017), ecological validity refers to how evidence should be sought, aligning methods, materials, and assessment situations with the natural world being examined.

For children with highly probable atypical development due to CZVS diagnosis, the home assessment expanded the possibility of expressing their development. The obstacles encountered from the first applications, resulting from considerable delays in different developmental domains, led us to the Bayley Manual to understand the adaptations and define conduct and procedures for fieldwork with this population. Workshops were conducted, and routines for monitoring and ongoing supervision were structured to ensure adaptation procedures for the instrument without affecting test integrity.

In the end, adapting the BSID-III to accommodate visual and motor difficulties made it possible to assess children's developmental function with CZVS accurately. There were 16 adaptations, organized according to the type of facilitation used (visual, motor, or general), applied in all assessment areas in this instrument. The observed data suggested the importance of constructing new perspectives in assessing children with atypical development, minimizing the interference of deficits (Araújo et al., 2017).

Final Considerations

It has been demonstrated that it is possible to conduct home assessments of the development of children with multiple disabilities in a population context, using adaptations provided by the Bayley-III Scales, with satisfactory levels of reliability. The team's training, supervision, and monitoring calibrated the assessors according to the reliability measures demonstrated by Kappa and ICC. Evidence of good performance of neuropsychological instruments in population studies investigating groups still underexplored in public health favors the research progress due to the results' credibility. It assists future researchers in choosing the tool.

In conclusion, the study contributes to advancing knowledge about children with Multiple Disabilities from the perspective of Public Health, providing reliability to the psychological assessment process in a population with multiple limitations in child development in the community context. Scientifically relevant investigations that examine findings for populations with complex developmental alterations are considered, emphasizing the need to test the accommodations required to maintain evidence of the validity and reliability of specific assessment tools.

References

Araújo, C. F., Cabral, C. B., Dantas, J., Oliveira, K. N. R. F., Flores, M. C. M., Almeida, T. M., Silva, T. C. L., Santos, L. M., & Santos, D. N. (2017). *Manual do Protocolo de Facilitações Sensoriais e Motoras para uso da Escala Bayley de Desenvolvimento Infantil – BSIDIII em crianças com Síndrome Congênita do Zika Vírus*. https://repositoriohml.ufba.br/handle/ri/31933

Bayley, N. (2006). *Manual of Bayley Scales of Infant Development TM*. (3rd ed.). The Psychological Corporation.

Blair, M.; & Hall, D. (2006). From health surveillance to health promotion: the changing focus in preventive children's services. *Archives of Diseases in Childhood*, *91*(9), 730–735. https://doi.org/10.1136/adc.2004.065003

Campos, D., Santos, D. C. C., Gonçalves, V. M. G., Goto, M. M. F., Arias, A.V., Brianeze, A. C. G. S., Campos, T. M., & Mello, B. B. A. (2006). Agreement between scales for screening and diagnosis of motor development at 6 months. *Jornal de Pediatria*, *82*(6), 470–474. https://doi.org/10.2223/JPED.1567

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or parcial credit. *Psycological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Diamond, A. (2000). Close interrelation of motor development and cognitive development and of the cerebelun and prefrontal cortex. *Child Developement*, *71*, 44–56. https://doi.org/10.1111/1467-8624.00117

Ferreira, R. C., Mello, R. R., & Silva, K. S. (2014). Neonatal sepsis as a risk factor for neurodevelopmental changes in preterm infants with very low birth weight. *Jornal de Pediatria*, *90*, 293–299. https://doi.org/10.1016/j.jped.2013.09.006

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, *5*(1), 105–112. https://doi.org/10.1177/014662168100500115

Gourladins, J. B., & Sá, C. S. C. (2022). *Desenvolvimento e saúde mental na infância*. Editora Ampla.

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence Based Nursing*, *18*(3), 66–67. doi.org/10.1136/eb-2015-102129

Hentges, C. R., Silveira, R. C., Procianoy, R. S., Carvalho, C. G., Filipouski, G. R., Fuentefria, R. N., Marquezotti, F., Terrazan, A. C. (2014). Association of late-onset neonatal sepsis with late neurodevelopment in the first two years of life of preterm infants with very low birth weight. *Jornal de Pediatria*, *90*, 50–57. https://doi.org/10.1016/j.jped.2013.10.002

Kettler, R. J. (2012). Testing Accommodations: theory and research to inform practice. *International Journal of Disability, Development and Education*, *59*(1), 53–66. https://doi.org/10.1080/1034912X.2012.654952

Madaschi, V., & Paula, C. S. (2011). Medidas de avaliação do desenvolvimento infantil: uma revisão de literatura nos últimos cinco anos. *Cadernos de Pós-Graduação em Distúrbios do Desenvolvimento*, *11*(1), 52–56. https://editorarevistas.mackenzie.br/index.php/cpgdd/article/view/11173

Madaschi, V., Mecca, T. P., Macedo, E. C., & Paula, C. S. (2016). Baykey III Scales of Infant and Toddler Development: Transcultural Adaptation and Psycometric Properties. *Paidéia*, *26*(64), 189–197. https://doi.org/10.1590/1982-43272664201606

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Ministério da Saúde. (2016). *Diretrizes de estimulação precoce: crianças de zero a 3 anos com atraso no desenvolvimento neuropsicomotor decorrente de microcefalia* [Early stimulation guidelines: children aged 0 to 3 years with delayed neuropsychomotor development due to microcephaly]. Secretaria de Atenção à Saúde. https://www.sbp.com.br/fileadmin/user_upload/2016/01/Diretrizes-de-Estimulacao-Precoce_Microcefalia.pdf.

Pasquali, L. (2017). Validade dos testes. *Revista Examen*, *1*(1), 14–48. https://examen.emnuvens.com.br/rev/article/view/19/17

Rodrigues, O. M. P. R. (2012). Escalas de desenvolvimento infantil e o uso com bebês. *Educar em Revista*, *43*, 81–100. https://doi.org/10.1590/S0104-40602012000100007

Ruiter, S. A., Nakken, H., van der Meulen, B. F., & Lunenborg, C. B. (2010). Low Motor Assessment: A Comparative Pilot Study with Young Children With and Without Motor Impairment. *Journal of developmental and physical disabilities*, *22*(1), 33–46. https://doi.org/10.1007/s10882-009-9165-5

Sandler, A. D., Brazdziunas, D., Cooley, C. W., De Pijem, L. G., Hirsch, D., Kastner, T. A., Kummer, M. E., Quint, R. D., Ruppert, E. S., Anderson, W. C., Crider, B., Burgan, P., Garner, C., McPherson, M., Michaud, L., Yeargin-Allspp, M., Cartwright, D., Johnson, C. P., & Smith, K. (2001). Developmental Surveillance and screening of infants and young children. *Pediatrics*, *108*, 192–195. https://doi.org/10.1542/peds.108.1.192

Santos, D. N., Araujo, T. M., Santos, L. M., Kuper, H., Aquino, R., Silveira, I. H., Miranda, S. S., Pereira, M., & Werneck, G. L. (2022). The Salvador Primary Care Longitudinal Study of Child Development (Cohort-DICa) Following the Zika Epidemic Study Protocol. *International journal of environmental research and public health*, *19*(5), 2514. https://doi.org/10.3390/ijerph19052514

Santos, D. C. C., & Ravanini, S. G. (2006). Aspectos do diagnóstico do desenvolvimento motor. In: Ribeiro, M.V., & Gonçalves, V. M., *Neurologia do desenvolvimento da criança*. (pp. 258–269). Revinter.

Souza, A. C., Alexandre, N. M. C., & Guirardello, E. B. (2017). Propriedades psicométricas na avaliação de instrumentos: avaliação da confiabilidade e da validade [Psychometric properties in instruments evaluation of reliability and validity]. *Epidemiologia e Serviços de Saúde (Brasília)*, *26*(3), 649–659. https://doi.org/10.5123/S1679-49742017000300022

Spooner, D. M., & Pachana, N. A. (2006). Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology*, *21*(4), 327.

Visser, L., Ruiter, S. A., Van der Meulen, B. F., Ruijssenaars, W. A., & Timmerman, M. E. (2013). Validity and suitability of the Bayley-III Low Motor/Vision version: A comparative study among young children with and without motor and/or visual impairments. *Research in developmental disabilities*, *34*(11), 3736–3745. https://doi.org/10.1016/j.ridd.2013.07.027

Visser, L., Ruiter, S. A., van der Meulen, B. F., Ruijssenaars, W. A., & Timmerman, M. E. (2014). Accommodating the Bayley-III for motor and/or visual impairment: A comparative pilot study. *Pediatric physical therapy: The official publication of the Section on Pediatrics of the American Physical Therapy Association*, *26*(1), 57–67. https://doi.org/10.1097/PEP.0000000000000004

Zwaigenbaun, L., & Penner, M. (2018). Autism spectrum disorder: Advances in diagnosis and evaluation. *BMJ*, *361*, 1674. https://doi.org/10.1136/bmj.k1674

Wheeler A. C. (2018). Development of Infants with Congenital Zika Syndrome: What Do We Know and What Can We Expect? *Pediatrics*, *141* (Suppl 2), 154–S160. https://doi.org/10.1542/peds.2017-2038D

Wheeler, A. C., Toth, D., Ridenour, T., Lima Nóbrega, L., Borba Firmino, R. Marques da Silva, C., Carvalho, P., Marques, D. Okoniewski, K., Ventura, L. O., Bailey, D. B., Jr, & Ventura, C. V. (2020). Developmental outcomes among young children with congenital Zika syndrome in Brazil. *JAMA Network Open;* *3*(5), e204096. https://doi.org/10.1001/jamanetworkopen.2020.409