



Boletín Médico del Hospital Infantil de México (English Edition)

www.elsevier.es/bmhim



REVIEW ARTICLE

Comparative analysis of child development screening tools designed and validated in Mexico

Rodrigo Orcajo-Castelán^{a,d,e,*}, Beatriz Sidonio-Aguayo^b,
Jorge Augusto Alcacio-Mendoza^c, Giovana Lucía López-Díaz^c

^a Unidad Médica de Alta Especialidad Hospital Regional de Psiquiatría Unidad Morelos, Instituto Mexicano del Seguro Social, México, D.F., México

^b Departamento de Rehabilitación Pediátrica, Hospital Infantil de México Federico Gómez, México, D.F., México

^c Facultad Mexicana de Medicina, Universidad La Salle, México, D.F., México

^d Principles and Practice of Clinical Research Course, Spaulding Rehabilitation Center, Massachusetts General Hospital, Harvard Medical School, Cambridge, Massachusetts, U.S.A.

^e Centre For Global Mental Health, King's College London/London School of Hygiene and Tropical Medicine, London, UK

Received 2 November 2015; accepted 3 November 2015

KEYWORDS

Child development;
Screening test;
Bias risk;
Validation;
Mexico

Abstract

Background: In recent years a number of child development screening tools have been developed in Mexico; however, their properties have not been compared. The objective of this review was to compare the report quality and risk bias of the screening tools developed and validated in Mexico in their published versions.

Methods: A search was conducted in databases, gray literature and cross references. The resultant tests were compared and analyzed using STARD, QUADAS and QUADAS-2 criteria.

Results: “VALoración NEuroconductual del DEsarrollo del LActante” (VANEDELA), “Evaluación del Desarrollo Infantil or EDI” (CDE in English), “Prueba de Tamiz del Neurodesarrollo Infantil” (PTNI), “Cartillas de Vigilancia para identificar alteraciones en el Desarrollo del Lactante” (CVDL) and “Indicadores de riesgo del Perfil de Conductas de Desarrollo” (INDIPCD-R) were included for comparison. No test fulfilled all STARD items. The most complete in their methodological description were VANEDELA and CDE. The areas lacking more data on the reports were recruiting and patient selection (VANEDELA, PTNI, CVDL, INDIPCD-R). In QUADAS evaluation, all had some risk bias, but some serious concerns of risk bias were raised by patient sampling and by the choice of the gold standard in two tests (PTNI, INDIPCD-R).

* Corresponding author.

E-mail: rodrigo_aquilino.orcajo_castelan@kcl.ac.uk (R. Orcajo-Castelán).

PALABRAS CLAVE

Neurodesarrollo;
Pruebas de tamiz;
Riesgo de sesgo;
Validación;
México

Conclusions: Child development screening tests created and validated in Mexico have variable report quality and risk bias. The test with the best validation report quality is VANEDELA and the one with the lowest risk bias is CDE.

© 2015 Hospital Infantil de México Federico Gómez. Published by Masson Doyma México S.A. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Análisis comparativo de pruebas de tamiz para la detección de problemas en el desarrollo diseñadas y validadas en México

Resumen

Introducción: En los últimos años se han desarrollado varias pruebas de tamiz para el desarrollo infantil temprano (DIT) en menores de 5 años en México. El objetivo de esta revisión fue comparar la calidad del reporte de validación publicado y riesgo de sesgo entre diferentes pruebas desarrolladas y validadas en México.

Métodos: Se realizó una búsqueda en bases de datos, literatura gris y referencia cruzada documental. Se efectuó un análisis comparativo de la calidad del reporte (STARD) y el riesgo de sesgo (QUADAS y QUADAS-2).

Resultados: Se incluyeron las siguientes cinco pruebas: Valoración Neuroconductual del Desarrollo del Lactante (VANEDELA), Evaluación del Desarrollo Infantil (EDI), Prueba de Tamiz del Neurodesarrollo Infantil (PTNI), Cartillas de Vigilancia para identificar alteraciones en el Desarrollo del Lactante (CVDL) e Indicadores de riesgo del Perfil de Conductas de Desarrollo (INDIPCD-R). Ninguna cumplió el 100% de los ítems de acuerdo con STARD. Las más completas en su descripción fueron VANEDELA y EDI. El área de procedimientos de muestreo fue en la que hubo menor cumplimiento (VANEDELA, PTNI, CVDL, INDIPCD-R). En QUADAS, todas las pruebas presentaron algún riesgo de sesgo. Las más importantes fueron la selección de la muestra y la elección del estándar de oro, que en dos estudios se identificó que no era el más adecuado (PTNI, INDIPCD-R).

Conclusiones: Las pruebas de tamiz mexicanas para el DIT varían en la calidad de reporte publicado y riesgo de sesgo. La de mejor calidad de reporte de validación es VANEDELA y la de menor riesgo de sesgo en los datos publicados es la prueba EDI.

© 2015 Hospital Infantil de México Federico Gómez. Publicado por Masson Doyma México S.A. Este es un artículo Open Access bajo la licencia CC BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Early identification of alterations in child development is essential to the well-being of children and their families because it allows providing an accurate diagnosis and implementing an early intervention for those having some type of alterations.¹ For the children who receive them, such interventions are associated with better adulthood functionality at multiple spheres² in addition to having a very high cost-benefit ratio.³

The comparison of clinical characteristics of each test is important to assist in the selection of the most appropriate tool for development evaluation. Moreover, it is essential to compare their quality of report and risk of bias because biased reports of diagnostic or screening tests results may end in the widespread adoption of tools that produce an inaccurate risk classification, thereby leading medical personnel to make an incorrect referral, diagnosis or treatment decision.⁴

A systematic review and comparative analysis of the literature done in 2012 by Romo-Pardo et al. found 13 screening tests that were created and validated in America for the timely identification of problems related to child develop-

ment, but none of them with data published for the Mexican population in scientific journals (except Denver-II).⁴ In recent years a significant number of screening tests have been created, some of which are already being applied to children <5 years of age in Mexico. This information on the validation and properties is not found in indexed journals, and a comparison has not been made between them.

Based on the above, the aim of this paper was to compare the quality of the validation reports published and their risk of bias among the screening tests developed and validated in Mexico.

2. Methods**2.1. Search and analysis**

Because part of the information about the tests created in Mexico is not available in search engines of scientific journals, during October 2015 an exhaustive search was carried out for child development screening tools in children <5 years of age, designed and validated in Mexico from 1980 to

date. A simple strategy was carried out using the Spanish terms “neurodesarrollo” or “desarrollo infantil” and “tamizaje” as well as the English terms “child development” and “screening”, in PubMed/MEDLINE, Scopus, Web Of Science, EMBASE, EBSCO, Google Scholar, LiLACS and SciELO, limiting the results to tests elaborated in Mexico.

Through different search strategies, cross references and asking experts in the field, seven screening tests created in Mexico for neurodevelopment evaluation were identified. Using the name of the tests, a new exhaustive search was performed looking for reports of their validation. The gray literature was screened in addition to the previously consulted sources. For the analysis only those where a published validation report was found were included, excluding the remaining reports.

2.2. Tools used for evaluation

2.2.1. STARD

The Standards for Reporting of Diagnostic Accuracy (STARD)^{5,6} were designed with the objective of improving the quality of the reports of diagnostic accuracy studies. They consist of a checklist of 25 items and flow diagrams that report the patient selection methods, the order in which the tests are conducted and the number of patients that should be evaluated with the index test and reference standard. They evaluate if the publications provide sufficient information that would allow detecting potential study bias and evaluate the potential generalization and applicability of the results.

2.2.2. QUADAS

The “Quality Assessment of Studies of Diagnostic Accuracy in Systematized Reviews” tool (QUADAS)⁷ was developed to assist in the evaluation of risk of bias in publications of diagnostic accuracy. It was elaborated based on three reviews of existing literature by using a Delphi methodology procedure, with the participation of a group of nine experts. It is composed of checklist of 14 qualitative items, which includes: patient spectrum characteristics of the sample, bias in disease progression, bias of verification, bias of review, bias of clinical review, bias of incorporation, how the test was done and how indeterminate results are managed. These items should be scored with the response “yes” if it is believed that the study being analyzed meets the characteristics described in each item; “no” if it does not meet the characteristics, or “unclear” if the test does not contain sufficient information to make a judgment.

2.2.3. QUADAS-2

The QUADAS-2⁸ tool was designed with the objective of evaluating the risk of bias in the studies of diagnostic accuracy. It is a structured questionnaire with open questions grouped in four domains, which include patient selection, index test, reference standard, and time elapsed between the index test and reference standard. The tool should be completed in four stages: the first intends to establish a review question; the second, to develop a specific guide for review; subsequently, evaluate the flow diagram published or, in case it has not been published, create one with the data provided and, finally, establish a judgment about bias and applicability. Each domain is evaluated in terms of *risk of bias*, and the first three are also evaluated in terms of

applicability concerns. To aid in establishing a decision on the risk of bias, some signal questions were included. This tool allows the evaluators to create a tabular presentation for each study evaluated, classifying each item as *low risk*, *high risk* or *unclear risk*.

2.3. Evaluation procedure

The analysis was performed in stages. In the first stage, each of the authors evaluated, separately and independently, the quality of the validation report using the STARD⁶ tool, and the risk of study bias with the QUADAS⁷ and QUADAS-2⁸ tools. The results of this evaluation, including flow diagrams and checklists, were collected using the formats developed for each tool. In the second stage, the evaluation formats from each of the authors were compared. Discrepancies were resolved by consensus. A final evaluation was produced, which was then transferred to the aforementioned formats and transformed for its graphic presentation.

3. Results

3.1. Identification of the screening tests

Seven screening tests of neurodevelopmental alterations created in Mexico were identified; Scale for Integral Child Development (Escala de Desarrollo Integral del Niño)⁹, Evaluation of the Neurodevelopment of the Newborn (Evaluación del Neurodesarrollo del Neonato, EVANENE),¹⁰ Neurobehavioral Evaluation of the Newborn (Valoración Neuroconductual del Desarrollo del Lactante, VANEDELA),¹¹ Child Development Evaluation (CDE) (Evaluación del Desarrollo Infantil, EDI),¹² Screening Test for Child Neurodevelopment (Prueba de Tamiz del Neurodesarrollo Infantil, PTNI),¹³ Surveillance Manuals to identify changes in Child Development (Cartillas de Vigilancia para identificar alteraciones en el Desarrollo del Lactante, CVDL)¹⁴ and Indicators of the Risk Profile of Behaviors in Development (Indicadores de Riesgo del Perfil de Conductas de Desarrollo, INDIPCD-R).¹⁵ All publications found using the scientific search engines were related with the CDE test¹⁶⁻¹⁸ with the exception of one related with NPED (Neuropediatric Development),¹⁹ which was excluded because it evaluated a developmental tool created in Cuba and had no concurrent design validation. The remainder of the tests were located in non-indexed scientific publications (INDIPCD-R¹⁵, CVDL²⁰), accessed through web pages (PTNI²¹), or as books, manuals, institutional research protocols or as masters/doctoral thesis (EVANENE²², VANEDELA²³, CDE¹⁶⁻¹⁸, PTNI²¹). Three had published validation articles in scientific journals, one had the validation data published online and another as a thesis report. Finally, five tests were included for evaluation. Excluded from this study were the Scale for Integral Child Development because no validation data were found, and EVANENE for the same reason because only a thesis of its validation was found, but doing so as an instrument for screening for brain damage.

3.2. General characteristics of the tests evaluated

The general characteristics of the tests evaluated are described in Table 1. A large variation in the age ranges eval-

Table 1 General description of the screening tests compared in the study.

Test	Age range	Modality of evaluation	Aspects evaluated	Scoring system	Criteria of normality and abnormality
Valoración Neuroconductual del desarrollo del Lactante (VANEDELA) ²³	1-24 months (evaluated in six groups)	Direct observation	Developmental behavior (feeding, reflexes, standing, gross motor, fine motor, eye-hand coordination, visual perception, auditory perception, receptive language, expressive language, cognition (emotional, social), developmental reactions (reflexes)) Warning or alert signs	Each question related to developmental behavior and developmental reactions has three possible responses: yes, no or not applicable. Warning or alert signs have four possible responses: absent, present or not recorded. The result is expressed as normal, doubtful, abnormal and not applicable (validation criteria if two items are marked as not applicable) Each question has two possible responses: yes or no. Results are based on a "traffic light" system: normal developmental (green), developmental lag (yellow) and risk of delay (red)	A score is obtained using 10 items for requested behaviors by age group, establishing as normal if all are marked, 8-9 as doubtful, abnormal if less than 7
CDE ¹⁶	1-59 m (evaluated in 14 groups)	Caregiver interview and direct observation	Areas of development (fine motor, gross motor, language, social and knowledge, biological risk factors, alert signs, alarm signs and neurological exploration)	Each item has two possible responses: yes or no (1 point given for each affirmative responses). Results are expressed in four categories: normal neurodevelopment, slight delay (low risk), moderate delay (moderate risk) and severe delay (high risk) Items are organized according to complexity from 1-64. The manual establishes the item number to be carried out for each given age. Results are expressed dichotomously as risk of delay or normal development	There are specific qualification criteria by age groups (1 m, 2-4 m and 5-59 m)
Prueba de Tamiz para Evaluar en Neurodesarrollo Infantil (PTNI) ²¹	11-49 m (evaluated in five groups)	Direct observation	Areas of development (gross motor, fine motor, language, cognition, socio-affective, independence)	Each item has two possible responses: yes or no (1 point given for each affirmative responses). Results are expressed in four categories: normal neurodevelopment, slight delay (low risk), moderate delay (moderate risk) and severe delay (high risk)	Score is obtained using the 12 behaviors requested in the items by age group behaviors by age group establishing cut-off points specifically for children 12-24 m and others for those 36-48 m
Cartillas de Vigilancia para identificar alteraciones en el Desarrollo del Lactante (CVDL) ²⁰	1-24 m (evaluated in 13 groups)	Caregiver interview	Not grouped by areas or domains	Items are organized according to complexity from 1-64. The manual establishes the item number to be carried out for each given age. Results are expressed dichotomously as risk of delay or normal development	Evaluates the child to perform the behaviors indicated by the manual and the previous age group, to classify as normal development
Indicadores de conductas de riesgo del Perfil de Conductas del Desarrollo (INDIPCD-R) ¹⁵	0-4 years (evaluated in six groups)	Caregiver interview	Areas of development (sitting, crawling, standing, walking, expressive language, receptive language, emotional/social, manual abilities, cognition and practice), antecedents	Each question is qualified using an ordinal scale, depending if one behavior is frequently observed (75-100%), sometimes (15-75%) or never (0-15%). Results are expressed as slight, moderate or severe developmental alteration	A score is obtained by adding indicators of risk to the child's chronological age. This score is compared to a standard scale obtained by SD: <1 SD, slight; 1-2 SD, moderate alteration; >2 SD, severe alteration

m, months; SD, standard deviation.

uated was found. CDE is the test that evaluates the broadest age range (1-59 months). Most of the tests use direct observation as the only modality for evaluation except for the CDE, which has a mixed evaluation modality: direct observation/caregiver structured interview. None of the tests uses paper-based home interview as an exclusive modality, e.g., filling out a questionnaire by the parents. The areas evaluated show a large variability although they generally adjust to the areas recommended in the literature: motor, language, adaptive or cognitive, personal or social.^{24,25}

Only INDIPAD-ETS and VANEDELA present the evaluation of neurological signs. The rating systems used differ widely but are adequately described. The same can be said of the criteria for abnormality except for the INDIPCD-R, whose published definition is confusing as it is unclear how the score obtained is compared with the gold standard. The properties of the screening tests reported in validation studies reviewed are summarized in Table 2.

The sample sizes used for the validations varied. The largest one was reported for the PTNI. The sample selection also varied as two tests were validated in populations selected from health institutions (VANEDELA and INDIPCD-R), two obtained their samples in specific populations (PTNI in rural and CVDL in urban) and only one presented a sample selection procedure intentionally balanced in terms of demographic characteristics and biological risk factors (CDE).

The gold standard used in the validation was also different for the different tests; three used a neurodiagnostic test such as the Gesell Developmental Schedules (GDS)²⁶ or Battelle-2 Developmental Inventory (BDI-2)²⁷ (VANEDELA, CDE, CVDL). One test used the diagnostic test from which (PCD-R) itself derives.^{28,29} Another test used as a proxy indicator a series of measures of the nutritional status, anemia and growth, alone and as composite scores (PTNI).

All tests reported sensitivity and specificity values as well as positive predictive values (PPV) and negative predictive values (NPV) that were adequate with what is recommended in the literature,²⁴ although there were tests that had a wide variation depending on the age group evaluated (VANEDELA) or the gold standard used (PTNI). In some age ranges they were too low to be used. Three tests did not describe the confidence intervals of their reported data. (VANEDELA, PTNI, CVDL).

3.3. Results of the evaluation of the quality of the report

Table 3 shows the checklist of the STARD tool. None of the tests evaluated complied with the entirety of the items to be reported. The most complete in their description were VANEDELA and CDE. The areas in which most tests were found with missing or incomplete information were those that refer to the description of the sampling procedure and selection of patients (VANEDELA, PTNI, CVDL, INDIPCD-R), the methods by which the missing data and the cases lost to follow-up were handled (All), description of the participants and the study flow diagram (PTNI, CVDL, INDIPCD-R, CDE) and the methods established for measuring the reproducibility of the test (INDIPCD-R).

3.4. Results of the evaluation of the risk of bias

Table 4 shows the results of the application of the tool for evaluation of the risk of bias QUADAS. None of the tests met 100% of the criteria evaluated. Only one test evaluated a spectrum of patients representative of the population (CDE). None reported management of the uninterpreted results, abandonment of the study or other case losses. Results of the QUADAS-2 evaluation are shown in Table 5 and are a final qualitative evaluation representing a consensus of the authors' opinion on how much risk of bias each tests presents. Because of its validation using a *proxy* gold standard, the PTNI presents a high risk of bias with respect to the standard of reference (weight for age, height for age, anemia and early stimulation not independent to the test).

The INDIPCD-R presents a high risk of bias of its index test and gold standard because it is validated against itself. Because they are samples by convenience without any type of adjustment, all tests have a high risk of bias with respect to patient selection. The one with the least risk was the CDE because it intentionally balanced all the groups evaluated.

4. Discussion

In previous reviews⁴ and reports,^{14,15} the comment about the few or no mentions of Mexican tests in the literature is repetitive as well as how difficult it is to find the validation of Latin American tests. The small number of results obtained when performing simple searches, limiting the results to Mexico, could be explained by two factors: the first would be the low level of visibility of Latin American journals because many are not indexed and therefore do not appear in search engines.³⁰ The second, due to a mix between a probable "fear" of international publications and a "malinchismo" effect towards national publications,³¹ which makes harder to take the decision to begin the laborious process of transforming the text of the thesis into scientific articles.³²

Because there is no ideal screening test for child development, it can be said that the general characteristics of all tests evaluated make them adequate for their use in Mexico. Before considering aspects on the quality of their validation, the decision to use one over the other should take into account its flexibility to be used.²⁴ In this regard, age ranges evaluated, variety in evaluation modalities and simplicity of the visual "traffic light" rating system favor the CDE test.

No study is found to be free of defects in the quality of its report. The most complete was the VANEDELA report, which is very broad and complete because it is published as a 180 page thesis. Although almost all data required by STARD for evaluating the reliability of the data are found, there are aspects of the methodology that limit its external validity such as the small sample size of each of the age groups evaluated and the process of sample recruitment, which was done by convenience in an urban population from clinical/hospital environments. Other validation reports such as the INDIPCD-R omit important data that may help to assess its validity, which makes its objective evaluation difficult. One possible solution for this phenomenon could be to extend the use of the evaluation tools mentioned in this article and use them as a guideline checklist to guarantee that the scientific papers are complete prior to publication.³³

Table 2 Comparison of the characteristics of the screening tests reported in the validation studies

Test (year of publication)	Study population	Gold standard and cut-off point for abnormality	Sensitivity	Specificity	PPV	NPV
VANEDELA (2001) ²³	184 children from an urban population who came to the neurodevelopmental follow-up laboratory at INP/UAM-X	TDQ <90 on GDS or at least one area of development evaluated on GDS with DDQ <76	1 m: 1.00 4 m: 0.88 8 m: 0.88 12 m: 0.96 18 m: 0.88 24 m: 0.88	1 m: 0.93 4 m: 0.75 8 m: 0.67 12 m: 0.45 18 m: 0.67 24 m: 0.78	1 m: 0.83 4 m: 0.91 8 m: 0.81 12 m: 0.69 18 m: 0.70 24 m: 0.78	1 m: 1.00 4 m: 0.69 8 m: 0.78 12 m: 0.90 18 m: 0.86 24 m: 0.88
Modified version of the CDE (2013) ^{16,18}	438 children in three Mexican states (202 children 16-60 months)	TDQ <90 on the BDI-2 ¹⁶	1-15 m: 0.74 CI (0.65-0.82)	1-15 m: 0.60 CI (0.51-0.68)	1-15 m: 0.61 CI (0.53-0.70)	1-15 m: 0.72 CI (0.63-0.81)
			16-60 m: 0.89 CI (0.82-0.95) Total 0.81 CI (0.64-0.86)	16-60 m: 0.62 CI (0.53-0.71) Total 0.61 CI (0.54-0.67)	16-60 m: 0.68 CI (0.60-0.76) Total 0.65 CI (0.59-0.70)	16-60 m: 0.86 CI (0.78-0.94) Total 0.78 CI (0.72-0.84)
		DDQ <80 on BDI-2 ¹⁸	Motor 0.84 CI (0.74-0.90) Communication 0.79 CI (0.71-0.85) Personal social 0.86 CI (0.77-0.92) Adaptive 0.91 CI (0.76-0.96) Cognitive 0.83 CI (0.76-0.88)	Motor 0.83 CI (0.87-0.90) Communication 0.79 CI (0.74-0.83) Personal social 0.85 CI (0.80-0.88) Adaptive 0.85 CI (0.81-0.88) Cognitive 0.88 CI (0.80-0.93)		

Table 2 (cont.)

PTNI (2013) ²¹	27,059 children in rural Mexican communities	Weight for age (Z-score ≤2)	12 m: 0.78	24 m: 0.73	12 m: 0.41	24 m: 0.52	12 m: 0.18	24 m: 0.22	12 m: 0.92	24 m: 0.91
		Height for age (Z-score ≤2)	12 m: 0.71	24 m: 0.62	12 m: 0.44	24 m: 0.58	12 m: 0.60	24 m: 0.71	12 m: 0.56	24 m: 0.48
		Anemia (cut-off point not specified)	12 m: 0.65	24 m: 0.60	12 m: 0.65	24 m: 0.60	12 m: 0.44	24 m: 0.28	12 m: 0.60	24 m: 0.78
		Stimulation (cut-off point not specified)	12 m: 0.92	24 m: 0.89	12 m: 0.60	24 m: 0.64	12 m: 0.50	24 m: 0.36	12 m: 0.94	24 m: 0.97
		All variables (cut-off point not specified nor combination of criteria)	12 m: 0.88	24 m: 0.86	12 m: 0.77	24 m: 0.79	12 m: 0.85	24 m: 0.74	12 m: 0.83	24 m: 0.91
CVDL (2014) ²⁰	2,702 children from urban areas of three Mexican states	Results < -2 SD in either of the scales of the two diagnostic tests: Bayley Development Scale II or on the GDS, "normal or abnormal diagnosis of the child"	0.84		0.75		0.83.		0.76	
INDIPCD-R (2015) ¹⁵	145 children coming from one clinic and two CENDI	PCD-R) (cut-off point not mentioned)	Clinic 1.00 CI (0.96-1.00) CENDI 0.94 CI (0.82-1.00)		Clinic 0.69 CI (0.40-0.98.17) CENDI 0.84 CI (0.61-1.00)		Clinic 0.90 CI (0.80-1.00) CENDI 0.90 CI (0.74-1.00)		Clinic 1.00 CI (0.94-1.00) CENDI 0.91 CI (0.71-1.00)	

VANEDELA, Valoración Neuroconductual del Desarrollo del Lactante; CDE, Child Development Evaluation; PTNI, Prueba de Tamiz del Neurodesarrollo Infantil; CVDL, Cartillas de Vigilancia para identificar alteraciones en el Desarrollo del Lactante; INDIPCD-R, Indicadores de riesgo del Perfil de Conductas de Desarrollo Revisado; PPV, positive predictive value; NPV, negative predictive value; m, months; INP, Instituto Nacional de Pediatría; UAM-X, Universidad Autónoma Metropolitana, Xochimilco campus; SD, standard deviation BDI-2, Battelle Developmental Inventory 2nd. Ed. (Spanish); TDQ, total development quotient; GDS, Gessell Developmental Schedules; DDQ, domain development quotient; PCD-R, Perfil de Conductas del Desarrollo; CENDI, Centro de Desarrollo Infantil.

Table 3 STARD checklist to evaluate the report quality of diagnostic precision studies

STARD checklist		VANEDELA ²³	CDE ¹⁶	PTNI ²¹	CVDL ²⁰	INDIPCD-R ¹⁵	
Section and theme	Item	Item description	Yes/incomplete/no; page number (p)				
Title/abstract/ key words Introduction	1	Identify the article as a study of diagnostic accuracy.	Yes; p. 1	Yes; p. 196	No	Incomplete	Yes; p. 135
	2	State the research questions or study aims.	Yes; p. 91	Yes; p. 197	No	Yes, p. 25	Yes; p. 138
Methods Participants	3	Describe the study population: the inclusion and exclusion criteria, setting and locations where data were collected.	Yes; pp. 91, 92	Yes; p. 197	Incomplete; p. 44	Incomplete; p. 26	Incomplete; p. 139
	4	Describe patient recruitment: What were the selection criteria used?	Yes; pp. 92, 97	Incomplete; p. 197	Yes; p. 44	No; p. 26	Yes; p. 138
	5	Describe patient sampling procedures	Yes; pp. 92, 97	Incomplete; p. 197	Incomplete; p. 44	Incomplete; p. 26	Yes; p. 138
	6	Describe data collection: Was data collection planned before the index test and reference standard were performed or after?	Yes; pp. 92, 97	Incomplete; p. 197	No	Incomplete; p. 26	Not mentioned
	7	Describe the reference standard and its rationale.	Yes; pp. 29-93	Yes; p. 197	Yes; p. 40	Yes; p. 26	Incomplete; p. 138
	8	Describe technical specifications of material and methods involved.	Yes; pp. 29-97	Yes; p. 197	Yes; pp. 46-48	Yes; p. 26	Yes; p. 139
Statistics	9	Describe definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard.	Yes; pp. 29-95	Yes; pp. 198, 201	Incomplete; pp. 41, 42	Yes; p. 26	Yes; pp. 138, 139
	10	Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.	Incomplete; p. 95	Yes; p. 197	Incomplete; p. 43	Incomplete; p. 26	Incomplete; p. 139
	11	Describe whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test.	No	Yes; p. 197	No	ND; p. 26	ND; p. 139
Statistics	12	Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty.	Incomplete; pp. 95-101	Yes; p. 197	No	No; p. 26	Yes; pp. 139-140
	13	Describe methods for calculating test reproducibility.	No	Incomplete; p. 197	Incomplete; p. 59	Incomplete; p. 26	No

Table 3 (cont.)

STARD checklist		VANEDELA ²³	CDE ¹⁶	PTNI ²¹	CVDL ²⁰	INDIPCD-R ¹⁵
Results	Participants					
	14	Report when study was done, including beginning and ending dates of recruitment	Yes; p. 97	No; p. 197	Incomplete; p. 44	No
	15	Report clinical and demographic characteristics of the study population: age, sex, spectrum of presented symptoms, etc.	Yes; pp. 98, 99	Yes; pp. 198, 200	Incomplete; pp. 45, 46	Yes; p. 140
	16	Report the number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard (A flow diagram is strongly recommended).	Incomplete; p. 100	No	Not reported	Incomplete; pp. 138, 139
	17	Report time interval from the index tests to the reference standard.	Yes; p. 92	Yes; p. 197	NR	No
	18	Report distribution of severity of disease (define criteria) in those with the target condition.	Incomplete; pp. 98, 99	Yes; pp. 197, 198	NR	Yes; p. 141
	19	Report a cross tabulation of the results of the index tests by the results of the reference standard.	Yes; pp. 100-133	Yes; p. 202	Yes; pp. 50, 51	Yes; p. 142
	20	Report any adverse events from performing the index tests or the reference standard.	NA	NA	NA	NA
	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty.	Incomplete	Yes	Incomplete; pp. 51, 52	Yes; p. 142
	22	Report how indeterminate results, missing responses and outliers of the index tests were handled.	Incomplete; p. 98	Incomplete	No	Incomplete; p. 139
Discussion	23	Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers.	NA	Yes; p. 203	No	Yes; p. 142
	24	Report estimates of the reproducibility of the test.	No	Incomplete; p. 197	Incomplete	No
	25	Discuss the clinical applicability of the study findings.	Yes; p. 143	Yes; p. 207	Yes	Yes; pp. 142-144

STARD, Standards for Reporting of Diagnostic Accuracy; VANEDELA, Valoración Neuroconductual del Desarrollo del Lactante; CDE, Child Development Evaluation; PTNI, Prueba de Tamiz del Neurodesarrollo Infantil; CVDL, Cartillas de Vigilancia para identificar alteraciones en el Desarrollo del Lactante; INDIPCD-R, Indicadores de riesgo del Perfil de Conductas de Desarrollo Revisado.

NR, not reported; NA, not applicable; ND, not determined.

Table 4 Questionnaire to evaluate risk bias in diagnostic accuracy studies (QUADAS)

Item	VANEDELA ²³	CDE ¹⁶	PTNI ²¹	CVDL ²⁰	INDIPCD-R ¹⁵
Was the spectrum of patients evaluated representative of those who will receive the test in practice?	No	Yes	No	No	No
Were selection criteria clearly described?	Yes	Yes	Yes	Uncertain	No
Is the reference standard likely to correctly classify the target condition?	Yes	Yes	No	Yes	No
Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	Yes	Yes	Uncertain	Yes	No
Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	Yes	Yes	Yes	Yes	No
Did patients receive the same reference standard regardless of the index test result?	Yes	Yes	Yes	Yes	Yes
Was the reference standard independent of the index test?	Yes	Yes	Yes	Yes	No
Was the execution of the index test described in sufficient detail to permit replication of the test?	Yes	Yes	Yes	Yes	Yes
Was the execution of the reference standard described in sufficient detail to permit its replication?	Yes	Yes	Uncertain	Uncertain	Yes
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes	Yes	No	Uncertain	Uncertain
Were the reference standard results interpreted without knowledge of the results of the index test?	Yes	Yes	No	Uncertain	Uncertain
Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	Yes	Yes	Yes	Yes	Yes
Were uninterpretable/ intermediate test results reported?	No	No	No	No	No
Were withdrawals from the study explained?	No	No	No	No	No

VANEDELA, Valoración Neuroconductual del Desarrollo del Lactante; CDE, Evaluación del Desarrollo Infantil; PTNI, Prueba de Tamiz del Neurodesarrollo Infantil; CVDL, Cartillas de Vigilancia para identificar alteraciones en el Desarrollo del Lactante; INDIPCD-R, Indicadores de riesgo del Perfil de Conductas de Desarrollo Revisado.

There is no bias-free scientific publication; however, there are procedures to reduce bias. The results of the two final evaluations (QUADAS and QUADAS-2) demonstrate that although the data of sensitivity and specificity are grossly similar among the tests compared, the validity of these data is compromised to varying degrees. Some of the publications evaluated omit basic data such as the measures of dispersion or precision of the measurement, and others

compromise the concurrent validation procedure by comparing the index test against itself or to a proxy measurement. Other tests such as VANEDELA also have a high risk of bias due to the small subsample size.

Similar to the conclusions in comparative reviews of developmental screening tests performed in other parts of the world,³⁴ it was found that among the tests for neurodevelopmental screening created in Mexico, none is perfect. The

Table 5 Consensus of the results of the application of evaluation tools at risk of bias (QUADAS-2)

Study	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
VANEDELA ²³							
CDE ¹⁶							
PTNI ²¹							
CVDL ²⁰							
INDIPCD-R ¹⁵							

Low Risk High Risk Unclear risk

VANEDELA, Valoración Neuroconductual del Desarrollo del Lactante; CDE, Child Development Evaluation; PTNI, Prueba de Tamiz del Neurodesarrollo Infantil; CVDL, Cartillas de Vigilancia para identificar alteraciones en el Desarrollo del Lactante; INDIPCD-R, Indicadores de riesgo del Perfil de Conductas de Desarrollo Revisado.

most flexible in its application and with less risk of bias in its validation results was the CDE test.

The tests for child development, created and validated in Mexico are, in general, adequate for their use in the local context although they have variable qualities of published reports and bias risks, and none is perfect. The test with the best validation report quality is VANEDELA followed by CDE, and the one with the lowest risk of bias in the data published is the CDE test. A comparative study of the screening tests vs. the gold standard will be required in order to establish which has the best properties.

Funding

No external funding was used for this study.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgments

We acknowledge the collaboration of Lic. Josué Laguna Hernández, Sección de Archivo Histórico, Biblioteca Dr. Ramón Villarreal Pérez, Universidad Autónoma Metropolitana-Xochimilco, in the localization of one of the tests evaluated in this study.

References

1. Council on Children with Disabilities; Section on Developmental Behavioral Pediatrics; Bright Futures Steering Committee; Medical Home Initiatives for Children with Special Needs Project Advisory Committee. Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics*. 2006;118:405-20.
2. Hamilton S. Screening for developmental delay: reliable, easy-to-use tools. *J Fam Pract*. 2006;55:415-22.
3. Glascoe FP, Foster EM, Wolraich ML. An economic analysis of developmental detection methods. *Pediatrics*. 1997;99:830-7.
4. Romo-Pardo B, Liendo-Vallejos S, Vargas-López G, Rizzoli-Córdoba A, Buenrostro-Márquez G. [Global neurodevelopmental screening tests for children under 5 years of age in the United States and Latin America: systematic review and comparative analysis]. *Bol Med Hosp Infant Mex*. 2012;69:450-62.
5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003;326:41-4.
6. Bossuyt P, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*. 2003;49:7-18.
7. Whiting PF, Rutjes AWS, Reitsma JB, Bossuyt PM, Kleijmen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003;3:25.
8. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529-36.
9. Atkin LC, Supervielle T, Sawyer R, Cantón P. Paso a paso: cómo evaluar el crecimiento y desarrollo de los niños. México D.F.: UNICEF, Editorial Pax; 1987.
10. Alvarado-Ruiz GA, Sánchez-Pérez MC, Mandujano-Valdés MA. EVANENE: Evaluación del Neurodesarrollo del Neonato. México D.F.: Universidad Autónoma Metropolitana Unidad Xochimilco; 2013.
11. Sánchez-Pérez MC, Benavides-González H, Mandujano-Valdés M, Rivera-González IR, Martínez-Vázquez RI, Alvarado-Ruiz GA. Valoración Neuroconductual del Desarrollo del Lactante (VANEDELA). México D.F.: Universidad Autónoma Metropolitana; 2007.
12. Comisión Nacional de Protección Social en Salud. Manual para la Aplicación de la Prueba Evaluación del Desarrollo Infantil "EDI". México, D.F.: Secretaría de Salud; 2013. 100p. Available from: <http://www.himfg.edu.mx/descargas/documentos/EDI/ManualparaPruebadeEvaluaciondelDesarrolloInfantil-EDI.pdf>
13. Ávila-Curiel CA, Álvarez-Izazaga MA, Lopez-Arce Coria AM, Medina-Carranza BE. Prueba de Tamiz del Neurodesarrollo Infantil. México, 2008. Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán. No. de registro 03-2008-120312060800.
14. Rivera R. Vigilancia del desarrollo en el primer nivel de atención. In: Rivera R, Sánchez C, editors. Vigilancia del desarrollo integral del niño. México: Editores de Textos Mexicanos; 2009. p. 9-15.
15. Bolaños C, Mina-Lara MM, Ramos-Ortega G. Indicadores de riesgo del Perfil de Conductas de Desarrollo como instrumento para la detección temprana de retrasos y alteraciones en el desarrollo. *Acta Pediatr Mex*. 2015;36:135-46.
16. Rizzoli-Córdoba A, Schnass-Arrieta L, Liendo-Vallejos S, Buenrostro-Márquez G, Romo-Pardo B, Carreón-García J, et al. Validación de un instrumento para la detección oportuna de problemas del desarrollo en menores de 5 años en México. *Bol Med Hosp Inf Mex*. 2013;70:195-208.
17. Rizzoli-Córdoba A, Ortega-Ríosvelasco F, Villasis-Keever MA, Pizarro-Castellanos M, Buenrostro-Márquez G, Aceves-Villagrán D, et al. Reliability for detection of developmental problems using the semaphore from the Child Development Evaluation test: Is a yellow result different from a red result? *Bol Med Hosp Infant Mex*. 2014;71:277-85.
18. Rizzoli-Córdoba A, Schnass-Arrieta L, Ortega-Ríosvelasco F, Rodríguez-Ortega E, Villasis-Keever MA, Aceves-Villagrán D, et al. Child Development Evaluation Test analysis by field improves detection of developmental problems in children. *Bol Med Hosp Infant Mex*. 2014;71:154-62.
19. Guadarrama-Celaya F, Otero-Ojeda GA, Pliego-Rivero FB, Porcayo-Mercado MR, Ricardo-Garcell J, Perez-Ábalo MC. Screening of neurodevelopmental delays in four communities of Mexico and Cuba. *Public Health Nurs*. 2012;29:105-15.
20. Rivera-González R, Villanueva-Romero Y, Amaro-López L, Sánchez-Pérez C, Figueroa-Olea M, Soler-Limón KM. Validez concurrente de las cartillas de vigilancia para identificar alteraciones en el desarrollo del lactante. *Ciencias Clínicas*. 2014;15:22-9.
21. Ávila-Curiel A, Álvarez-Izazaga MA, Reidl-Martínez L, López-Arce AM. Vigilancia Epidemiológica del Neurodesarrollo Infantil en Comunidades Marginales de México. Capítulo III Elaboración de la prueba de tamiz del neurodesarrollo infantil (PTNI). Available from: <http://www.nutricionenmexico.org.mx/encuestas/neuro%20web.pdf>
22. Palma-Reyes CR. Valor diagnóstico de dos instrumentos clínicos en la detección de daño cerebral perinatal en prematuros. Informe preliminar (Tesis de Licenciatura). Universidad Autónoma de Querétaro. Querétaro, México; 2014.
23. Martínez-Vázquez RI. Evaluación de la sensibilidad y especificidad del instrumento de detección: Valoración Neuroconductual del Desarrollo del Lactante. (VANEDELA) (Tesis de Maestría). Universidad Autónoma Metropolitana-Xochimilco. México D.F., México; 2001.

24. Rydz D, Shevell MI, Majnemer A, Oskoui M. Developmental screening. *J Child Neurol*. 2005;20:4-21.
25. Shevell M, Majnemer A, Platt RW, Webster R, Birnbaum R. Developmental and functional outcomes in children with global developmental delay or developmental language impairment. *Dev Med Child Neurol*. 2005;47:678-83.
26. Ball RS. The Gesell developmental schedules: Arnold Gesell (1880-1961). *J Abnorm Child Psych*. 1977;5:233-9.
27. Newborg J. Development and standardization. In: Newborg J, editor. *Battelle Developmental Inventory*. Itaska, IL: Riverside Publishing; 2004. pp. 95-148.
28. Bolaños C. *Manual del Perfil de Conductas de Desarrollo Revisado*. México D.F.: Instituto de Terapia Ocupacional; 2005.
29. Bolaños C. Estudio de validación del perfil de conductas de desarrollo (PCD) como instrumento para evaluar el nivel de desarrollo de los niños de 0-4 años. *Bol Med Hosp Infant Mex*. 2003;60:79-89.
30. Romero-Torres M, Acosta-Moreno LA, Tejada-Gómez MA. Ranking de revistas científicas en Latinoamérica mediante el índice *h*: estudio de caso Colombia. *Rev Esp Document Cient*. 2013;36:e003. doi: <http://dx.doi.org/10.3989/redc.2013.1.876>
31. Cicero-Sabido R. ¿Por qué los autores mexicanos no envían trabajos importantes a las revistas médicas mexicanas? Un comentario breve. *Gac Med Mex*. 2006;142:128-9.
32. Gross Portney L, Watkins MP. Reporting the results of clinical research. In: *Foundations of Clinical Research. Applications to Practice*. New York: Pearson Prentice Hall; 2008. pp. 879-900.
33. Cook C, Cleland J, Huijbregts P. Creation and critique of studies of diagnostic accuracy: use of the STARD and QUADAS methodological quality assessment tools. *J Man Manip Ther*. 2007;15:93-102.
34. Bedford H, Walton S, Ahn J; Policy Research Unit in the Health of Children, Young People and Families; Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health. *Measures of Child Development: A Review*. London: University College London (UCL); 2013. 103p. Available from: https://www.ucl.ac.uk/cpru/documents/review_of_measures_of_child_development