PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2013, 50(2), 147-162

Pontificia Universidad Católica de Chile
Facultad de Educación

Development and Standardization of a Tablet-based Assessment Tool of Early Childhood Education Learning Goals

Construcción y estandarización de un instrumento de evaluación de aprendizajes esperados en educación parvularia basado en tablet

[1]Doyna Illmer, [2]Ricardo Rosas, [2]Soledad Véliz, [2]María Paz Ramírez, [2]Andrés Aparicio, [2]Catalina Benavente, and [2]Carolina Thibaut

[1]Department of Education, Pontificia Universidad Católica de Chile
[2]Centro de Desarrollo de Tecnologías de Inclusión (CEDETi UC), Chile

Abstract

In Chile there are not enough reliable, standardized instruments to assess young children's knowledge of basic concepts defined essential by the Ministry of Education of Chile. The goal of this research is to validate a play-based assessment tool of basic concepts for young children (3 to 6 years old) called *dip*, supported in tablet. The instrument assesses basic concepts related to spoken language and mathematical reasoning. It was standardized with 360 children between 3 and 6 years old attending institutions offering early childhood education in the Metropolitan Region of Chile. The results show good internal reliability (0.7) and consistent reliability for different groups of assessed knowledge. There is evidence of strong correlation with a computer-based assessment instrument of early literacy competencies (0.7) and a moderate correlation with a test of basic concepts in a traditional format (0.5).

Keywords: preschool education, assessment, learning goals, instrument development

Post to:
Doyna Illmer Craciun
Department of Education, Pontificia Universidad Católica de Chile
Av. Vicuña Mackenna 4860, Macul, Santiago, Chile,
Email: dkillmer@uc.cl
Research conducted in the context of FONDEF project D09I1238 "Construcción y estandarización de instrumentos de evaluación psicométrica para el desarrollo de una cultura de evaluación ética y rigurosa [Construction and standardization of psychometric assessment tools for the development of an ethical and rigorous assessment culture]." The theoretical development was supported by developments financed by the Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) under Project 111095.
This article was originally written in Spanish.

ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi:10.7764/PEL.50.2.2013.9

Resumen

En Chile escasean los instrumentos de evaluación de aprendizajes para párvulos que cuenten con criterios de confiabilidad y validez y que evalúen aprendizajes esenciales definidos de ese modo por el Ministerio de Educación de Chile. El objetivo de esta investigación es validar una prueba de evaluación de aprendizajes a través del juego llamada *dip,* sostenida en *tablet*. La prueba evalúa conceptos básicos relacionados con lenguaje verbal y razonamiento lógico-matemático en niños de entre tres y seis años. La prueba *dip* fue estandarizada con 360 niños entre tres y seis años de instituciones de diversa dependencia administrativa que ofrecen educación parvularia en la Región Metropolitana de Chile. La prueba conlleva evidencia de buena confiabilidad interna (0,7) y confiabilidad consistente por eje de aprendizajes evaluados. Asimismo, evidencia una alta correlación con una prueba de evaluación de competencias lectoras iniciales (0,7) y una correlación moderada con una prueba de aprendizajes en formato tradicional (0,5).

Palabras clave: educación parvularia, evaluación, aprendizajes esperados, construcción de instrumento

In Chile, the assessment of expected learning outcomes in early childhood education is experiencing a surge in interest, aimed at guiding evidence-based educational decisions to improve this level of education. *Dip*, a formative assessment tool, provides educators with information about children's achievements in relation to the expected outcomes for logical-mathematical reasoning and quantification and for spoken language. This instrument allows the classroom teacher to make evidence-based adjustments to the proposed curriculum, in order to address diversity of learning (RDA – *respuesta a la diversidad de aprendizaje*), identify at-risk children in a timely manner, and refer children to specialists, as appropriate.

In Chile, early childhood educators have few tools to support decisions made in classroom teaching. This is important since it is one of the low performance areas of educators according to the country's teacher performance assessment system (*Sistema de Evaluación del Desempeño Profesional Docente*), which was established in 2003 and achieved national coverage by 2005. This system primarily assesses quality of education based on criteria established in the framework for good teaching (CPEIP & MINEDUC, 2008). According to recent data, early childhood educators receive the lowest scores on this assessment, lower than primary and secondary education teachers (Manzi, González, & Sun, 2011). Within the areas evaluated, the following aspects are rated as basic or unsatisfactory: (a) contribution of the activities to class learning goals (78%); (b) attribution of learning outcomes (77%); (c) ability to reshape teaching strategies (72%); and (d) analysis of pedagogical decisions (68%).

These four areas assessed as basic or unsatisfactory are essential to achieving the proposed expected learning outcomes for early childhood education. *Dip* yields easily interpretable reports that are divided into different dimensions: low, in process, expected, or higher than expected, which clearly identify the aspects that need to be reinforced or strengthened, as well as areas where children have already achieved the expected learning outcomes and should be exposed to new, challenging yet attainable, learning. The *dip* results report includes details on the items that the child answered incorrectly, enabling the educator to accurately modify the curriculum coverage to achieve the expected learning outcomes according to the Early Childhood Education Curriculum Foundations (BCEP - *Bases Curriculares de la Educación Parvularia*). Thus, *dip* provides specific and detailed information in a user-friendly and intuitive format that allows the professional in charge of the children's learning to make evidence-based instructional decisions, to reinforce, strengthen, challenge, intervene, or refer the child to a specialist, as appropriate.

What is *dip*?

Dip is designed to assess, in a formative and complementary way, the expected learning outcomes of children aged 3 to 6 in early childhood education. *Dip* evaluates two core areas of learning: logical-mathematical relationships and quantification, and spoken language. To operationalize the expected learning outcomes described in the BCEP, *dip* is based on the learning goals described in the Early Childhood Education Progress Maps (MPEP - *Mapas de Progreso de la Educación Parvularia*), an operational systematization of the BCEP.

Theoretical basis and references for the creation of *dip*

Child learning and development

Development and learning are closely related. This has important implications for teaching, since identifying children's levels of development, whether in relation to themselves or others, allow teachers to support their learning (Bransford, Derry, Berliner, Hammerness, & Beckett, 2005). To do this, educators need to understand how children learn (Bransford et al., 2005) and the general developmental progressions and individual differences of development. Knowing how instruction can support development is one of the central characteristics of effective teaching (Bransford et al., 2005).

Early childhood education in Chile

In the first six years of life, children are particularly susceptible to the development of basic skills, particularly those associated with optimal performance in formal education (Leseman, 2002). There is consensus on the importance of education in early childhood. When receiving a quality preschool education, children develop the knowledge, skills, and dispositions necessary for learning in kindergarten and beyond. This increases the likelihood of continuing positive achievement patterns, and reduces the chances of academic failure and dropout. Investing in early childhood education is efficient and effective for economic and workforce development. The sooner the investment is made, the greater the return (Heckman & Masterov, 2004).

Children do not have an innate preparedness to successfully meet the demands of formal education. In this context, school readiness can be understood as:

> Processes that change as children acquire important school skills in the first few years of schooling. This includes not only the level of skill at school entry (…) but also the slope or growth in those skills over time, as a function of child and family characteristics as well as of classroom characteristics and school context. (Vernon-Feagans, Odom, Pancsofar, & Kainz, 2007, p. 63).

This definition of school readiness emphasizes the need to incorporate both the contexts and relationships that affect child development (Knoche, Sheridan, Edwards, & Osborn, 2010), integrating the many variables that influence the skills a child has when beginning formal education. The fit between the child, family, school, and the community leads to developmental processes that change over time (Vernon-Feagans et al., 2007).

While academic skills in early childhood development tend to receive more credit as predictors of later academic performance (Kolb & Fantie, 2009), recent research in this field has also placed great importance on the development of executive functions (Welsh, Nix, Blair, Bierman, & Nelson, 2010). These functions develop substantially between the ages of 3 and 5 (Welsh et al., 2010) and, conceptually, are the functions necessary for children to organize their thinking and behavior with "increasing flexibility (...) and to engage in self-regulated and rule-governed behavior" (Welsh et al., 2010, p. 2). Some of these functions, such as working memory and self-control, can facilitate school readiness and early childhood learning (Welsh et al., 2010).

Basic relational concepts

Basic relational concepts are essential tools of thought used in all cultures (Siegler, 1998) and are crucial for school readiness.

Dip assesses concepts in the areas of logical-mathematical reasoning and quantification and spoken language. According to Boehm (2001), being familiar with, understanding, and knowing how to use these concepts allow children to understand and describe relationships between objects and the location and characteristics of people, places, and things; understand the order of events, follow instructions from the educator/teacher and standardized tests; become involved in emergent literacy activities; and participate in problem-solving activities that include classification, comparing, sequencing, and identification of different attributes.

As children move towards elementary education, knowledge of these basic concepts become increasingly critical to understanding what is being communicated in the classroom (Boehm, 2013), making it an important element to assess in the early childhood education years.

Early Childhood Education Curriculum Foundations (BCEP)

The BCEP were prepared by the Chilean Ministry of Education's Curriculum and Evaluation Unit. Their implementation began in 2002 and they were published in 2005. They were made to upgrade, redirect, and enrich the learning contexts and opportunities offered to children, and to integrate and coordinate a single body of expected learning outcomes and educational guidelines into one curricular instrument with common criteria (MINEDUC, 2005).

The curricular foundations are organized by interrelated dimensions of learning experiences, and they organize expected learning outcomes for all children from the first months of life until 6 years of age or entry to elementary education. Every dimension of learning experience is composed of core learning areas plus the respective general learning goal. Likewise, the expected learning outcomes are organized into two learning levels. The first level comprises the first months of life to 3 years of age, and the second level comprises 3 to 6 years of age, or when children begin primary education.

Early Childhood Education Progress Maps (MPEP - Mapas de Progreso de la Educación Parvularia)

The MPEP were developed in 2004 by the Chilean Ministry of Education, along with the National Board of Nursery Schools (JUNJI - *Junta Nacional de Jardines Infantiles*) and the Fundación Integra, with the participation of UNICEF. The Progress Maps can be understood as the operationalization of the Curriculum Foundations. There are differences between the BCEP and MPEP, particularly that the maps are directed at the level or section in which the child is located in relation to his or her learning achievements.

Learning Achievement Profiles in Early Childhood Education (PLAEP-R - Perfiles de Logro de Aprendizaje en la Educación Parvularia)

PLAEP-R (2009), developed by Fundación Integra and Pontificia Universidad Católica de Chile, served as a reference in the creation of *dip*. This test was piloted nationally between 2008 and 2009. It is a traditional-format standardized test for the educational community that aims to improve the educational process in nursery schools and early childhood education centers by measuring the learning achievement level of children between 1 and 5 years old. As it was created in accordance with the Early Childhood Education Curriculum Foundations, it provides comparable information with respect to the learning areas defined therein.

The total sample (PLAEP-R) for standardization consisted of 1192 children attending Fundación Integra schools, JUNJI schools, private centers, and municipal, charter, and private schools throughout the country. PLAEP-R and *dip* are tests that, while both using the curriculum foundations to determine what to assess, differ in item format, portability, standardization of items, time, automation of the correction process, and correction time. Also, *dip* is a type of formative assessment, allowing the classroom teacher to make decisions in a timely manner to make improvements before the end of the school year.

Instrument description

Dip was developed based on an assessment through play model. Playing is not only a privileged learning experience (Rosas et al., 2003); it is also "a window to a child's level of development" (Kelly-Vance & Ryalls, 2005, p. 549). Assessment through play has a high ecological validity, and the National Association of School Psychologists (NASP) of the United States identifies it as an appropriate way to assess young children's needs (Kelly-Vance & Ryalls, 2005). *Dip* is tablet-based. The assessment was designed as a digital game with a narrative where the main character captures the attention and stimulates

the motivation of children during the tasks requested of them. The main character of *dip*, Caja, or "Box," is gender-neutral, so that both boys and girls can identify. Box goes through several stages in order to find a lost pet. The game's narrative intended to maintain *flow*, a state of optimal experience where a person is so involved in the activity that his or her self-awareness and sense of time disappear (Shute & Kim, in press). This psychological state helps the assessment not be perceived as such by the students being assessed.

What does *dip* evaluate?

Dip assesses a set of expected learning outcomes that reflect two dimensions of the BCEP: the *relationship with the natural and cultural environment* and *communication*, as well as the core learning areas *logical-mathematical relationships and quantification* and *spoken language* associated with each dimension. There is at least one item per core learning area.

In addition to the expected learning outcomes in accordance with the Ministry of Education curriculum guidelines, *dip* also assesses reading precursors associated with phonological awareness such as sound recognition and first and last syllables. Finally, *dip* evaluates basic relational concepts essential for children to follow directions, understand the content of instructional material, engage in activities inside and outside the classroom, and communicate effectively with others.

Thus *dip* provides a formative and complementary alternative to natural observation, one of a kind in Chile, for assessing learning achievements in early childhood education. These data enable evidence-based pedagogical decisions, which positively contribute to improving the quality of education.

Instrument features and structure

Dip is played in two formats, both on the tablet: Mundos, or "Worlds," (60 items) and Entremundos, or "Between Worlds," (32 items), with a total of 92 items. It takes about 20 to 30 minutes to complete. Correction is automatic, except for two manual correction items (emergent writing and oral communication).

The Worlds format is a continuously moving background or scene (like a conveyor belt) where items are presented to the child for a short period. The response time of the Worlds is the same for all items, except for some that require the item to stop so the child can observe the images in detail and then answer. The movement resumes once the child answers. The items that have not been answered in the Worlds appear again at the end of each scene, so that the child has another chance to respond. Meanwhile, the items belonging to Between Worlds are problems presented on a static screen and assess learning in a format where the item can be answered in a practically unlimited amount of time. For a sample item, see Figure 1.

In each *dip* item, the evaluated subject's response is followed by the occurrence of a stimulus (in the form of an animation and/or sound). This stimulus does not tell the child if his or her answer was right or wrong, but rather keeps the game dynamic.

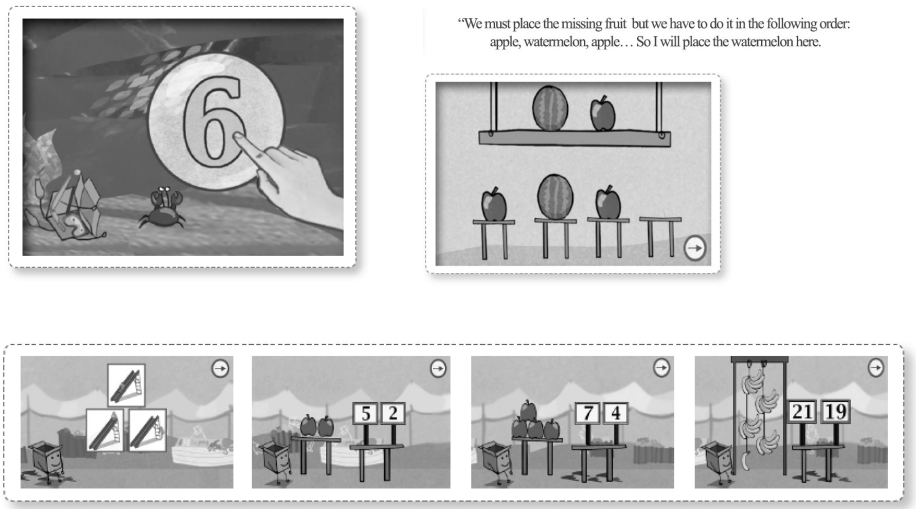

Figure 1. Sample items.

Method

Participants

The sample population consisted of children attending early childhood education centers in the Metropolitan Region of Chile. A convenience and intentional sample was created through contact with the UC Department of Education, finding a similar number of institutions with different administrative units: schools pertaining to JUNJI, Fundación Integra, and the Ministry of Education, and private institutions. Ultimately there was a list of 30 institutions offering early childhood education for children between 3 and 6 years of age. The central criteria for segmenting the sample were administrative unit (municipal, charter, and private) and age (three age ranges). Care was taken to maintain a balance in terms of gender. The study only included children who submitted consent forms signed by their responsible caregivers. Children with sensory and/or severe motor disabilities were excluded.

In the absence of family income information to determine socioeconomic status (traditionally associated with learning outcomes and child development), the school's administrative unit was used as an approach to the socioeconomic status of the families of tested children[1].

The sample for the instrument standardization consisted of 360 children from institutions with different administrative units (see Table 1). Less than 10% of the sample was lost during testing and later analysis, and the analyses were conducted with 333 children.

Table 1
Characteristics of the experimental sample

Admin. Unit	Sex	3 yr – 3 yr 11 m	4 yr – 4 yr 11 m	5 yr – 5 yr 11 m	6 yr – 6 yr 11 m	Total
Municipal	Boy	18	17	16	18	69
	Girl	18	16	15	15	64
Charter	Boy	10	17	15	13	55
	Girl	8	8	18	16	50
Private	Boy	13	16	17	13	59
	Girl	12	17	16	18	63
	Total	79	91	97	93	360

[1] Among OECD countries, Chile has the lowest levels of social inclusion in its schools where less than 50% of the variance in socioeconomic status lies in these institutions (OECD, 2012). Therefore, it is unlikely to find children from different socioeconomic backgrounds enrolled in the same schools.

Method for creating the items

Dip was developed in three phases: pre-pilot, pilot, and experimental application. Based on the Chilean Ministry of Education guidelines on expected learning outcomes for Chilean children, and together with a group of experts in education and developmental psychology, concepts were defined that could be evaluated with tablet touch technology and that were age-appropriate. Early versions of the test had four age ranges: from 3 years 0 months to 3 years and 6 months; from 3 years and 6 months to 3 years and 11 months; from 4 years to 5 years and 11 months; and from 6 years to 6 years and 11 months. The pre-pilot phase took place in August 2011. Items were applied to a sample of six children in order to test the functionality of the platform, the reaction times to the items, playability, etc. After this, programming and design errors were resolved. The pilot took place in October 2011, using a test version with 28 items per age group. In total, 60 children from institutions with three types of administrative units (private, charter, and municipal) were assessed. At the time of the pilot evaluation, there were three items, of increasing difficulty, for each expected learning outcome. All the children evaluated, regardless of age, were exposed to all the items in order to generate the levels of difficulty of each item. The total test duration was about 60 minutes.

The items to be included in the experimental phase were reviewed by a panel of experts, and according to the changes made in the pre-pilot and pilot phases, the items that discriminated by level of difficulty were selected. That is, items that in the pilot phase were answered equally well or equally poorly by all age groups were eliminated. About a quarter of the items were removed. All expected learning outcomes ended up being represented by at least one item (since each had initially been represented by three items). For the experimental phase only two items, of varying levels of difficulty, were used per expected learning outcome. The total test duration was approximately 20 to 30 minutes. For details of the number of items by core learning areas tested, see Table 2. For details of the procedure according to the phase of the study, see Table 3.

Table 2
Dip items of and their relationship to learning dimensions and core areas

	Dimensions of learning experiences											
	Personal and social education			Communication				Relationship with the natural and cultural environment				
Core learning areas	Autonomy	Identity	Coexistence	Spoken language			Artistic languages	Living things and their environment	Human groups, their ways of life, and significant events	Logical-mathematical relationships and quantification		**TOTAL**
Progress Maps (3 to 4 years)				OC	ER	EW				LMR	Q	
dip	0	0	0	4	22	3	0	0	0	42	21	**92**

Table 3
Details of the procedure by study phase

	March 2011	October 2011	November-January 2012
Phase	Pre-pilot application	Pilot application	Experimental application
Criteria to assess	1. Assessment of difficulty by age group 2. Assessment of grouping of items into categories (quantification, understanding, time and space)	Application to sample of 60 children from three types of administrative units	Application to sample of 360 children from three types of administrative units
Changes	1. Item reduction 2. Ensure progression of difficulty	1. Correction of design and programming errors 2. "Ceiling effect" detected in the assessment of all ages 3. Items are modified 4. Assessment guidelines are modified 5. It is ensured that each evaluator has a tablet	
Characteristics of evaluators		1. Eight with degrees in Early Childhood Education	1. Ten with degrees in Early Childhood Education and Psychology
Number of items	39 per age group	28 items per age group plus 12 Between Worlds	92 items in total, for all ages
Test time	60 minutes per age group	45 minutes total	20 to 30 minutes total
Quality criteria	1. All items apply to all children of all ages	1. Online correction platform implemented 2. Training session held for evaluators; application manual distributed 3. Informed consent given to guardians	1. Evaluators are trained in two sessions 2. Focus group held with evaluators halfway through application 3. Correction platform gives random cases to evaluators for score assignment 4. Cross correction is used: if two evaluators disagree on the score for the same case, it goes to a third evaluator who acts as a judge 5. The educational institution provides a space for assessment during the school day

Application procedure

For the experimental sample, schools were first approached by phone, and then a meeting was arranged with the directors of the institution. Educational centers considered eligible were those who formally joined the study through their directors. In each of the selected institutions, informed consent was given to the children's guardians via the students. Only children who returned the consent form signed by a responsible caregiver were included in the assessment. The selection criteria for evaluators administering *dip* in nursery schools and schools were: degree in early childhood education with completed internships and classroom experience working with children, and, if possible, recommendations. Evaluators with a degree in psychology and four or more years of experience in fieldwork or clinical practice were also included. All evaluators submitted their academic and professional résumés and participated in three

training sessions of two hours each. In these sessions, evaluators were trained in the proper use of the tablet and how to give the instructions for different items and subtests. The sessions also specified proper presentation and behavior in educational institutions, defined the evaluators' responsibilities and commitments, covered basic knowledge of assessment and the assessment context, and resolved any doubts. Following the training, evaluators were selected according to their on-site performance, time availability, commitment, and responsibility.

One of the study's limitations, with implications for *dip*'s validation, is that the sample is not representative of the country, being concentrated in the metropolitan region.

Results

The means and standard deviations indicate that there are gender differences, where girls perform better than boys in all areas of learning (see Table 4 for means by sex). These differences are statistically significant, especially in logical-mathematical reasoning F (1.331) = 11.88, p = 0.001 and quantification F (1.290) = 8.35, p = 0.004 (see Table 5 for the statistically significant differences by sex).

Table 4
Means by gender and learning area

Sex		*dip* LMR	*dip* Q	*dip* OC	*dip* ER	*dip* EW
Female	Mean	26.7	14.4	2.3	15.9	3.0743
	N	163	163	163	163	148
	Stan. dev.	7.8	4.5	0.8	4.0	1.9
Male	Mean	23.7	13.3	2.1	14.8	2.4
	N	170	170	170	170	144
	Stan. dev.	7.8	4.9	1.0	4.3	2.0

Note: LMR = logical-mathematical relationships, Q = quantification, OC = oral communication, ER = emergent reading, EW = emergent writing.

Table 5
Differences by gender and learning area

			Sum of squares	df	Quadratic mean	F	Sig.
dip Logical-mathematical reasoning * Sex	Intergroup	(combined)	728.9	1	728.9	11.88	0.001**
	Intragroup		20307.1	331	61.3		
	Total		21036.1	332			
dip Quantification * Sex	Intergroup	(combined)	97.8	1	97.8	4.34	0.038*
	Intragroup		7451.0	331	22.5		
	Total		7548.8	332			
dip Oral communication * Sex	Intergroup	(combined)	4.3	1	4.3	4.86	0.028*
	Intragroup		294.2	331	.889		
	Total		298.5	332			
dip Emergent reading * Sex	Intergroup	(combined)	96.1	1	96.1	5.42	0.021*
	Intragroup		5872.6	331	17.7		
	Total		5968.8	332			
dip Emergent writing * Sex	Intergroup	(combined)	32.9	1	32.9	8.35	0.004**
	Intragroup		1142.8	290	3.9		
	Total		1175.7	291			

Differences were observed by administrative unit and learning area. Municipal establishments had lower total raw scores on *dip* than charter and private schools (M = 51, SD = 12.3). These differences were repeated in all areas of learning (see Table 6 for means by administrative unit). Further analysis indicated that the differences between schools were not statistically significant for oral communication F (2.33) = 0.59, p = 0.551 and introduction to writing F (2.28) = 0.70, p = 0.496 (see Table 7 for significant differences by administrative unit and learning area).

Table 6
Means by administrative unit

Type of school		*dip* Score	*dip* LMR	*dip* Q	*dip* OC	*dip* ER	*dip* EW
	Mean	51.2	23.1	12.5	2.2	14.4	2.5
Municipal	N	131	131	131	131	131	111
	Stan. dev.	12.3	7.4	4.0	.9	3.6	1.9
	Mean	57.8	26.1	14.7	2.2	16.3	2.8
Charter	N	96	96	96	96	96	86
	Stan. dev.	14.9	8.4	5.4	1.0	4.4	2
	Mean	57.9	26.8	14.8	2.3491	15.7	2.8
Private	N	106	106	106	106	106	95
	Stan. dev.	13.9	7.6	4.6	.90546	4.4	2.1

Note: LMR = logical-mathematical relationships, Q = quantification, OC = oral communication, ER = emergent reading, EW = emergent writing.

Table 7
Significant differences by administrative unit and learning area

			Sum of squares	df	Quadratic mean	F	Sig.
	Intergroup	(combined)	3575.4	2	1787.7	9.62	0.000**
dip score * admin. unit	Intragroup		61307.7	330	185.7		
	Total		64883.2	332			
	Intergroup	(combined)	912.5	2	456.2	7.48	0.001**
dip LMR * admin. unit	Intragroup		20123.5	330	60.9		
	Total		21036.1	332			
	Intergroup	(combined)	418.8	2	209.4	9.69	0.000**
dip Q * admin. unit	Intragroup		7129.9	330	21.6		
	Total		7548.8	332			
	Intergroup	(combined)	1.0	2	0.5	0.59	0.551
dip OC * admin. unit	Intragroup		297.5	330	0.9		
	Total		298.5	332			
	Intergroup	(combined)	212.4	2	106.2	6.08	0.003*
dip ER * admin. unit	Intragroup		5756.3	330	17.4		
	Total		5968.8	332			
	Intergroup	(combined)	5.6	2	2.8	0.70	0.496
dip EW * admin. unit	Intragroup		1170.0	289	4.0		
	Total		1175.7	291			

Note: LMR = logical-mathematical relationships, Q = quantification, OC = oral communication, ER = emergent reading, EW = emergent writing.

Evidence of reliability

Dip assesses expected learning outcomes grouped in five areas (logical-mathematical reasoning, quantification, oral communication, emergent writing, emergent reading). Cronbach's alpha for *dip* total scores is 0.7.

The quantification area consists of 21 items (α = 0.8), the logical-mathematical reasoning area consists of 42 items (α = 0.9), the oral communication area consists of 4 items (α = 0.8), and the emergent reading area consists of 22 items (α = 0.7). With respect to the emergent writing area, it should be noted that the responses of children aged 3 and 4 are assessed based on structure and content criteria, while children aged 5 and 6 are evaluated based on structure, content, and length criteria. Since the criteria for each age group are different, for the reliability analysis the children were divided into two age groups: emergent writing area for children aged 3 and 4 has two items (α = 0.9) and for children aged 5 and 6 there are three items (α = 0.9).

Evidence of reliability for each learning area is detailed in the following table:

Table 8
Evidence of reliability by learning area

	Cronbrach's alpha	No. of items
Quantification area	0.8	21
Logical-mathematical reasoning area	0.9	42
Oral communication area	0.8	4
Emergent reading area	0.7	22
Emergent writing area		
3 and 4 years old	0.9	3
5 and 6 years old	0.9	

Evidence of validity

Progression of *dip* test scores by age of the sample

When comparing children's average results on *dip* and in the different areas, one can see a growing and significant progression based on the age of the children ($p < 0.001$). This is expected based on the learning that children achieve throughout development, which makes this an instrument that can discriminate between different ages. The following table shows the means and standard deviations for *dip* scores and each of the learning areas for the different age groups (see Table 9 and Figure 2).

Table 9
Progression of average scores by age

Age	3 years	4 years	5 years	6 years
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Quantification	17.8 (5.7)	24.1 (6.2)	29.3 (6.4)	32.3 (5.6)
Logical-mathematical reasoning	6.7 (2.3)	9.2 (2.7)	11.7 (2.8)	13.5 (2.8)
Oral communication	3.6 (2.7)	6.7 (2.9)	8.3 (3.0)	10.3 (3.2)
Emergent reading	9.0 (2.8)	9.8 (2.9)	11.6 (3.0)	14.0 (2.9)
Emergent writing	0.8 (1.1)	3.3 (1.8)	3.3 (1.8)	3.8 (1.9)

Evidence of concurrent validity

The results obtained by the sample on *dip* were correlated with other tests that assess similar constructs. In this case, *dip* performance was compared with two different tests: a) ABCDeti (*Prueba de Habilidades Lectoras Iniciales* [Initial Reading Skills Test], Rosas et al., 2011) and b) PLAEP-R (*Perfil de Logros de Aprendizajes en la Educación Parvularia* [Learning Achievement Profiles in Early Childhood Education],

revised version, Fundación Integra, 2009). In addition, to evaluate the effect of the technological platform as a potential invalidity factor in rural populations, a tablet and paper version of *dip* was administered to a sample of 26 children from a rural school.

Table 10 presents the results of the correlation between the total *dip* test scores and total ABCDeti test scores. These show that the test provides evidence of validity for assessing reading precursors (r = 0.7, n = 12, p = 0.01).

Table 10
Evidence of concurrent validity between *dip* and an initial reading skills assessment test

		dip	ABCDeti
dip	Pearson Correlation	1	0.7*
	Significance (bilateral)		0.01
	N	333	12
ABCDeti	Pearson Correlation	0.7*	1
	Significance (bilateral)	0.01	
	N	12	14

*Correlation is significant at level 0.05 (bilateral).

PLAEP-R (2009) is a traditional-format assessment tool whose standardized battery is only available for children of the Fundación Integra and is used to support the foundation in its regional and institutional decision-making. PLAEP-R was standardized for children between 1 and 5 years of age attending Integra establishments nationwide. This tool presents evidence of reliability, but no evidence of validity with an external criterion, in part because until 2012 it was the only instrument that existed in Chile to assess expected learning outcomes (Fundación Integra, 2009). Table 11 presents the results of the correlation between *dip* and PLAEP-R.

Table 11
Evidence of concurrent validity between *dip* and PLAEP-R

		Total PLAEP-R_Raw	Total *dip*-raw	*dip* score quantification	*dip* score oral communication	*dip* score logical-mathematical reasoning	*dip* score emergent reading	*dip* score emergent writing
Total PLAEP-R_Raw	Pearson correlation	1	0.520**	0.121	0.582**	0.591**	0.011	0.052
	Sig. (bilateral)		0.001	0.465	0.000	0.000	0.947	0.752
	N	39	39	39	39	39	39	39

**Correlation is significant at level 0.01 (bilateral)

As shown in Table 11, there is a significant, although moderate, correlation between the total scores of the scales (r = 0.520, n = 39, p = 0.001). This result gives preliminary evidence of concurrent validity. The moderate correlation can be explained by looking at the relationship between *dip* learning areas and total scores of the PLAEP-R; a moderate and highly significant relationship is seen in oral communication (r = 0.582, n = 39, p = 0.000) and logical-mathematical reasoning (r = 0.591, n = 39, p = 0.000), but not in quantification (r = 0.121, n = 39, p = 0.465), emergent reading (r = 0.011, n = 39, p = 0.947), or emergent writing (r = 0.052, n = 39, p = 0.752). Thus, although derived from the same concepts, *dip* and PLAEP-R assess separate aspects of expected learning outcomes in the aforementioned areas. While more research is needed to verify this, it is proposed that *dip* assesses initial reading skills that are not assessed in

PLAEP-R, and therefore acts as a complement to the spoken language area of PLAEP-R. With respect to emergent writing, further research is required to determine the validity of this scale. It is concluded that the tests do not share the aspect of quantification based on an item-by-item review of both tests, so it is more appropriate to consider the relationship as evidence of discriminant validity, which demonstrates the specificity of *dip*.

It is necessary, in future research, to compare the tests with a third criterion, in order to determine the nature of the assessment in the quantification and emergent writing scales.

Comparison of paper and digital *dip*

To rule out possible effects of invalidity due to the test format, a micro-study was conducted comparing performance in both digital and paper formats of a sample of 26 children from a rural school. The first method consisted of the administration of the test in its standard format, while the second was an adaptation of the test, administered in a printed format. Items that, by their nature, were not possible to adapt to pen and paper format were excluded (those that indicate motion, for example).

First, children's performance on the different aspects of the test was compared (see Table 12). As shown in the table, none of the differences was statistically significant, leading to the conclusion that the digital format is perfectly usable in populations that supposedly have a disadvantage due to lack of experience with digital platforms.

Table 12
Comparison of digital *dip* and paper *dip* scores

Areas	Digital	Analog	Significance
Logical-mathematical reasoning	17.8	18.8	NS
Quantification	8.2	8.1	NS
Oral communication	2.1	2.4	NS
Emergent reading	2.1	2.6	NS
Emergent writing	4.9	5.3	NS

Second, the correlation in performance between the two versions was studied, and the results suggested that the two versions are parallel, if one takes the correlation measure as a measure of reliability (see Table 13). The only scale that must be treated with caution when applied to populations with a technology disadvantage is the writing scale, which shows a reliability within the acceptable limit.

Table 13
Correlation of digital *dip* and paper *dip* scores

Areas	Correlation coefficient	Significance (bilateral)	N
Logical-mathematical reasoning	0.9	0.0	26
Quantification	0.8	0.0	26
Oral communication	0.9	0.0	26
Emergent reading	0.7	0.0	26
Emergent writing	0.8	0.0	26

Discussion

How *dip* can contribute to the teaching and learning process

The overall purpose of the assessment in an educational context is to help teachers and other stakeholders involved in education make evidence-based instructional decisions to inform and/or adjust the teaching and learning process (Brassard & Boehm, 2007).

Ideally, assessment is a dynamic and continuous process that uses a variety of measures and approaches, focuses on the child's learning context, is used to discover the child's strengths and emerging areas of development, supplies appropriate strategies and interventions, and is conducted from the perspective that the child will change. It also assumes that the earlier the intervention is made, the more likely it is to produce beneficial results (Brassard & Boehm, 2007). Lastly, parents need to be involved in the assessment process in many ways, not only to provide information about the child's development and specific needs, but also to foster a greater awareness of the importance of their presence in the child's development (Brassard & Boehm, 2007).

In Chile, early childhood education assessment has been linked to the measurement of development, intelligence, and readiness for admission to primary education. Most assessments are conducted in nursery schools and daycare centers (for example, PLAEP-R). These standardized tests are used to support decision-making at the national and institutional level, and the results (on the level of learning of children at the end of a process or educational level) enable the monitoring of progress and larger-scale issues. For a detailed comparison between *dip* and PLAEP-R, see Table 14.

Table 14
Comparison *dip* and PLAEP-R

Characteristics	*dip*	PLAEP-R
Type of assessment	With emphasis on formative assessment	Summative institutional-level measurement (results)
	Invisible assessment through play	Traditional assessment
Means of assessment	Touch screen digital device	Real objects
Assessed area	Two key learning areas: Logical-mathematical relationships and quantification and spoken language	All. At least one item per core learning area
Ages	3 to 6 years old	11 months to 5 years 3 months old
Portability	High; digital tablet	Limited; battery of about 15 objects per age
Standardization of items	Battery is standardized in terms of components (software loaded on a tablet)	Battery is made up of objects that must be bought
Total test time	20 to 30 minutes	Between 35 and 50 minutes depending on age
Type of correction	Almost fully automated to reduce scoring errors. Only two manual correction items	Manual, done at the time the child responds
Correction time	The items that require manual correction take five minutes	Around 10 minutes per item
Access to the test	For sale/by agreement	Limited unless belonging to Integra

As indicated above, a child's *dip* test results consist of the five areas explored throughout the test: logical-mathematical reasoning, quantification, oral communication, emergent reading, and emergent writing. Each child obtains a performance level for each area that may correspond to the expected level for his or her age or may show learning that is in process or requires intervention. All these performance levels are geared to criteria, and in particular, to the expected learning outcomes for the age of each child.

For *dip* report to be beneficial in the planning and the support that educators and other professionals provide to children, the disaggregated performance level is also specified by the relational concepts assessed.

As *dip* is a formative test, and considering the issues discussed above, if a child's score seems low, it is important to use complementary modes of assessment to obtain a more complete picture.

Perhaps *dip*'s most significant contribution, which contributes directly to improving the quality of early childhood education, is that it provides measurable learning goals and inputs to make evidence-based instructional decisions. The results allow the classroom teacher to propose specific plans to change curriculum coverage in order to respond appropriately to the diversity of learning in the classroom and/or to refer to a specialist when appropriate.

The challenges that remain for *dip*, which are being addressed by the authors in other current and future research projects, include: the need to conduct studies to examine the viability of *dip* for monitoring expected learning outcomes two and three times per year; the need to conduct research using *dip* as a tool to guide pedagogical decision-making in conjunction with instruments to assess executive functions, in order to provide more specific information for classroom work; and the need to conduct studies with a larger sample size, representative of the whole country, to improve the validity of the test. Future studies should especially consider the use of traditional variables associated with child learning and development, such as such as family income, to determine socioeconomic status, and gender. In this study the differences reported regarding gender and administrative unit are inconclusive, due to the aforementioned limiting representativeness. However, they merit deeper analysis. Finally, it is necessary to study a way to assess social skills and the ability to relate and interact positively with others, using this technology and the principle of invisible assessment.

The original article was received on June 3rd, 2013
The revised article was received on August 13th, 2013
The article was accepted on September 3rd, 2013

References

Boehm, A. (2001). *Boehm-3 Preschool: Boehm test of basic concepts*. Bloomington, MN: Pearson.

Boehm, A. (2013). *Assessment to intervention using the Boehm Test of Basic Concepts-Third Edition (Boehm-3),* 1-10. Retrieved from http://images.pearsonclinical.com/images/Assets/Boehm-3/Boehm_White_paper_RPT.pdf

Bransford, J., Derry, S., Berliner, D., Hammerness, K., & Beckett, K. (2005). Theories of learning and their roles in teaching. In L. Darling-Hammond, & J. Bransford (Eds.), *Preparing teachers for a changing world: what teachers should learn and be able to do* (pp. 40-87). San Francisco CA: John Wiley & Sons.

Brassard, M., & Boehm, A. (2007). *Preschool assessment: principles and practices*. New York, NY: The Guilford Press.

Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas (CPEIP), Ministerio de Educación (Mineduc) (2008). *Marco para la buena enseñanza*. Retrieved from http://www.docentemas.cl/docs/MBE2008.pdf

Fundación Integra (2009). *Perfil de logros de aprendizajes en la educación parvularia versión revisada (PLAEP-R).* Santiago, Chile: Fundación Integra.

Heckman, J., & Masterov, D. (2004). *The productivity argument for investing in young children.* [Working paper 5]. Retrieved from http://jenni.uchicago.edu/Invest/FILES/dugger_2004-12-02_dvm.pdf

Kelly-Vance, L., & Ryalls, B. O. (2005). Best practices in play assessment and intervention. In J. Grimes, & A. Thomas (Eds.), *Best practices in school psychology V* (pp. 549-560). Bethesda, MD: National Association of School Psychologists.

Knoche, L., Sheridan, S., Edwards, C., & Osborn, A. (2010). Implementation of a relationship-based school readiness intervention: a multidimensional approach to fidelity measurement for early childhood. *Early Childhood Research Quarterly*, *25*, 299-313. doi: 10.1016/j.ecresq.2009.05.003

Kolb, B., & Fantie, B. (2009). Development of the child's brain and behavior. In C. Reynolds, & E. Fletcher-Janzen (Eds.), *Handbook of Clinical Child Neuropsychology* (pp. 19-46). New York, NY: Springer Science + Business Media. doi: 10.1007/978-0-387-78867-8_2

Leseman, P. (2002). *Early childhood and care for children from low-income or minority backgrounds*. OECD: París.

Manzi, J., González, R., & Sun, Y. (2011). *La evaluación docente en Chile*. Santiago, Chile: Centro de Medición MIDE UC.

Ministerio de Educación (Mineduc) (2005). *Bases curriculares de la educación parvularia*. Santiago: Mineduc.

OECD (2012). *Panorama de la educación 2011: Indicadores de la OCDE*. Spain: Santillana. doi: 10.1787/9789264180482-es

Rosas, R., Medina, L., Meneses, A., Guajardo, A., Cuchacovich, S., & Escobar, P. (2011). Construcción y validación de una prueba de evaluación de competencia lectora inicial basada en computador. *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana, 48*(1), 43-61. doi: 10.7764/PEL.48.1.2011.4

Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., Flores, P., Grau, V., Lagos, F., López, X., López, V., Rodríguez, P., & Salinas, M. (2003). Beyond Nintendo: design and assessment of educational video games for first and second grade students. *Computers and Education*, *40*(1), 71-94. doi: 10.1016/S0360-1315(02)00099-4

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology (4th Edition)* (pp. 311-321). New York NY: Lawrence Erlbaum Associates, Taylor & Francis Group. doi: 10.1007/978-1-4614-3185-5_25

Siegler, R. (1998). *Cognition in children.* Upper Saddle River: Prentice Hall.

Vernon-Feagans, L., Odom, E., Pancsofar, N., & Kainz, K. (2007). Farkas and Hibel, a transactional/ecological model of readiness and inequality: person/process/context interactions. In A. Booth, & A. C. Crouter (Eds.), *Disparities in school readiness. How families contribute to transitions into school* (pp. 61-78). Mahwah, NJ: Erlbaum.

Welsh, J., Nix, R., Blair, C., Bierman, K., & Nelson, K. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology*, *102*(1), 43-53. doi: 10.1037/a0016738