PSYKHE
2021, 30(1), 1-15
https://doi.org/10.7764/psykhe.2018.22337

ISSN 0717-0297
www.psykhe.cl

Propiedades Psicométricas de un Test de Comprensión Lectora para Alumnos de Educación Primaria

Psychometric Properties of a Reading Comprehension Test for Primary School Students

Gema Pascual[1], Edurne Goikoetxea[2] y Héctor Bustos[3]

[1] Facultad Técnica, Universidad Católica de Temuco, Chile
[2] Facultad de Psicología y Educación, Universidad de Deusto
[3] Sede Villarrica, Pontificia Universidad Católica de Chile

Se presentan las propiedades psicométricas del test de Comprensión Lectora para Intervenir en Primaria (CLIP), desarrollado en base a modelos teóricos de los procesos cognitivos de la comprensión de textos (Kintsch, 1998). El objetivo del test es evaluar la comprensión lectora en estudiantes de primaria y arrojar información útil para planear la enseñanza que debe seguir a la evaluación. Con este propósito se usan diferentes textos (narrativos y expositivos, cortos y largos) y 3 tipos de preguntas (locales, globales e inferenciales), según la clasificación de Magliano et al. (2007), cuya respuesta debiera exigir al lector los procesos básicos de comprensión para construir el texto base y el modelo situacional. Por medio de un muestreo por conveniencia, participaron 1958 estudiantes de 3º a 6º de educación básica de 10 colegios particulares subvencionados y municipales de las comunas de Villarrica y Loncoche de la región de la Araucanía de Chile. El análisis de la consistencia interna (Kuder-Richardson 20) de las puntuaciones arroja valores iguales o superiores a 0,79. El análisis factorial confirmatorio halla evidencia de la presencia de 3 factores: Procesamiento Local, Procesamiento Global y Procesamiento Inferencial, cuando se analizan los resultados de los textos cortos. En los textos largos, el ajuste del modelo a los datos fue insatisfactorio. El análisis del funcionamiento diferencial de los ítems, según sexo y nivel socioeconómico, no reveló sesgo en ningún ítem del test.

Palabras clave: comprensión lectora, propiedades psicométricas, modelo de construcción-integración, procesos cognitivos, educación básica

This article presents the psychometric properties of the test of Reading Comprehension for Intervention in Primary School (CLIP), which was developed based on theoretical models of cognitive processes from current research on text comprehension (Kintsch, 1998). The purpose of the test is to assess reading comprehension in elementary school students and provide useful information to plan the teaching that should follow the evaluation. To this end, several texts are used (narrative and expository, short and long), along with 3 types of questions (local, global, and inferential), based on the classification by Magliano et al. (2007). To answer, readers should need to use basic comprehension processes to construct the basic text and situational model. Through a convenience sampling process, 1958 3rd to 6th grade students were recruited from 10 subsidized private and public schools located in the municipalities of Villarrica and Loncoche (Araucanía region of Chile). The internal consistency analysis (Kuder-Richardson 20) of CLIP scores yielded values equal to or greater than 0.79. The confirmatory factor analysis performed finds evidence of the presence of 3 factors in the comprehension of short texts: Local Processing, Global Processing, and Inferential Processing. In long texts, model fit was unsatisfactory. Differential item functioning analysis revealed no sex or socioeconomic bias in any of the items.

Keywords: reading comprehension, psychometric properties, construction-integration model, cognitive processes, elementary school

La evaluación de la comprensión lectora, entendiendo por comprensión el proceso mental de extraer y construir significado de la lectura atenta de un texto (Snow, 2002), puede servir a propósitos docentes, como

Gema Pascual https://orcid.org/0000-0002-9270-9898

Edurne Goikoetxea https://orcid.org/0000-0003-0922-5567

Este artículo fue parcialmente financiado por el proyecto HU2010-39 y la ayuda IT955-16 del Departamento de Educación, Universidades e Investigación del Gobierno Vasco.

La creación del CLIP no hubiera sido posible sin la colaboración de la Pontificia Universidad Católica de Chile Sede Villarrica y, muy especialmente, del decisivo apoyo recibido de su exdirector, Fernando Burrows, así como de la profesora Zoila Araneda Reyes.

La correspondencia relativa a este artículo debe ser dirigida a Gema Pascual, Facultad Técnica, Universidad Católica de Temuco, Manuel Montt, 56, Temuco, Araucanía, Chile. E-mail: gpascual@uct.cl. O también puede ser dirigida a Edurne Goikoetxea, Facultad de Psicología y Educación, Universidad de Deusto, Apartado 1, 48080 Bilbao, España. E-mail: egoikoetxea@deusto.es

conocer diferencias individuales, ayudar a entender el rendimiento escolar y planear una mejor intervención. La práctica de evaluar la comprensión en países como Chile y España ha ido en alza, gracias a la participación en programas de evaluación internacional, como PISA (siglas de Programme for International Student Assessment de la Organisation for Economic Cooperation and Development [OECD], (OECD, 2016, 2019; Schleicher & Tamassia, 2000) y nacional, como SIMCE (siglas de Sistema de Medición de la Calidad de la Educación de Chile; Agencia de Calidad de la Educación, 2013). Sin embargo, en el aula es infrecuente el uso de test estandarizados que ofrezcan información basada en la investigación sobre la comprensión de textos (Martínez et al., 2008; Snow, 2002) o que exijan pensamiento profundo y reflexión para ser realizados con éxito (Alexander & The Disciplined Reading and Learning Research Laboratory, 2012; O'Reilly et al., 2018). Además, apenas existen test de comprensión que se construyan alineados con programas o metodologías de intervención, en aras de medir el efecto de la intervención y su mantenimiento en el tiempo (O'Reilly et al., 2014; Scott, 2009), por lo que rara vez ofrecen información que ayude al maestro a decidir cómo y cuánto intervenir. En este trabajo, los autores están interesados en la creación y uso de test que se conviertan en un medio instructivo que facilite la consciencia y la acción de los profesores sobre las necesidades de enseñanza de la comprensión de textos.

Por otro lado, la investigación ha mostrado que los niños que saben leer en 3° de primaria no alcanzan automáticamente una adecuada comprensión lectora en cursos sucesivos (Snow, 2002; Spencer & Wagner, 2018) y, aunque la mayoría de los bachilleres norteamericanos examinados con test estandarizados dominan las habilidades para comprender textos simples, solo un 8% es capaz de hacer inferencias complejas (Ingels et al., 2005) y, al cabo de diez años, solo el 1% de los que tenían una puntuación en lectura comprensiva en el cuartil más bajo de la distribución lograron una titulación de máster y el 13% de bachiller, mientras que el 13% de los que tenían una puntuación en el cuartil superior lograron un máster y el 47% el título de bachiller (Lauff et al., 2015). Se sabe también que mientras más tarda un niño en adquirir el conocimiento, el interés y las habilidades de descifrado necesarias para comprender los textos según lo esperado para su edad, más dificultades afronta en el futuro (Alexander, 2003). Por esto, la detección temprana del bajo rendimiento lector y la enseñanza explícita de las estrategias cognitivas involucradas en la comprensión son hoy recomendadas desde informes (Snow, 2002) y metaanálisis (Kispal, 2008; National Reading Panel, 2000; Spencer & Wagner, 2018). Una recomendación de estos informes es fundar la creación de test en modelos teóricos y en resultados de la investigación sobre la comprensión. La finalidad del presente trabajo es precisamente construir un test llamado Test de Comprensión Lectora para Intervenir en Primaria (en adelante CLIP), basado en uno de los modelos psicológicos sobre los procesos cognitivos de la comprensión de textos.

Para construir el CLIP se consultaron, además de las orientaciones de los estándares para la construcción de test psicológicos y educativos ofrecidos por la American Educational Research Association, la American Psychological Association y el National Council on Measurement in Education (Standards for educational and psychological testing, 2014), el modelo de construcción-integración (Kintsch, 1998) y la investigación sobre buenos y malos comprendedores. El modelo de Kintsch supone, que durante la lectura, el lector construye tres niveles de representación del texto, en forma de proposiciones, aunque admite, sobre todo para el nivel más elaborado, el modelo mental de imagen de Johnson-Laird (1983): (a) superficial, esto es, el nivel más básico, el significado de las palabras y las frases precisas permanece poco tiempo en la memoria, no exige hacer inferencias; (b) texto base, esto es, el nivel de representación de las ideas del texto, construidas desde el análisis de la representación superficial en proposiciones, permanece de manera algo más estable en la memoria y (c) modelo situacional o nivel más elaborado, que es independiente de la estructura del texto e incorpora información del conocimiento previo y la experiencia del lector, además de sus objetivos y motivaciones en la lectura, y la elaboración de inferencias lingüísticas y elaborativas, permanece más tiempo en la memoria. Para Kintsch, la comprensión es justamente este nivel, el cual permite que el lector retenga la información por períodos largos de tiempo, la recupere con facilidad, genere nueva información y la utilice de manera divergente o en situaciones que requieren solución de problemas. Cuan elaborado o productivo sea el modelo situacional dependerá de la capacidad del lector para realizar inferencias, de sus conocimientos previos, de los propósitos de la tarea y del conocimiento de la estructura de los textos.

Por otra parte, dadas las limitaciones de memoria del lector, la comprensión es cíclica, lo que subraya la naturaleza dinámica e interactiva del proceso. Este modelo sugiere dos fases. Durante la fase de construcción, se activa el significado de las palabras, se forman las proposiciones y se generan las inferencias necesarias para mantener la coherencia local (relación entre frases sucesivas), lo cual no depende del contexto. Durante la fase de integración, se realizan inferencias necesarias para lograr un modelo coherente y global (las frases del texto tienen sentido en conjunto), lo cual sí depende del contexto. Además, en ambas fases actúan

componentes metacognitivos y motivacionales. De hecho, estudios actuales dan la misma relevancia a los factores cognitivos que a los factores afectivos (Afflerbach et al., 2013). La teoría de Kintsch no está exenta de limitaciones y aspectos no desarrollados (de Vega et al., 1999), pero ofrece hoy, hasta que lleguen futuros avances, un punto de partida para abordar un constructo como el de la comprensión de textos que no es unitario sino complejo y multifacético. Particularmente útil para el diseño del CLIP es la noción de niveles de representación del texto.

La investigación sobre buenos y malos comprendedores señala que el mal comprendedor no posee un perfil homogéneo y puede tener varias fuentes de dificultad, como son: (a) problemas con las habilidades para recuperar el significado a nivel de palabra o de frase (ayuda necesaria para construir la representación superficial de cada frase) y (b) problemas con los procesos para establecer la coherencia local y global de un texto (esto es, la capacidad de integrar los significados de las frases y de construir un modelo situacional del texto). La mayoría de estos estudios ha usado diseños correlacionales, pero, de los que han utilizado diseños para conocer la dirección de causa y efecto, se desprende que, en general, es en las habilidades para hacer inferencias y monitorear la comprensión donde puede estar el origen de las dificultades de comprensión (Cain, 2010). Por esta razón, el CLIP se basa en responder a preguntas que exigen hacer inferencias de distintos niveles, desde más locales a más globales.

Además, entre los test estandarizados actuales para medir comprensión lectora en lengua española, no abundan los que se basan en teorías explícitas y aceptadas y se centran en la medida de las habilidades para inferir (ver Martínez et al., 2008, para los test más usados; Montanero Fernández, 2004, para los enfoques y las técnicas para evaluar la comprensión; Snow, 2002, para una revisión en inglés). La ausencia de una teoría de base es una de las razones de lo difícil que resulta encontrar medidas válidas de la comprensión, que estén midiendo las habilidades deseadas y no otras solapadas (e.g., fluidez lectora), de las que es importante distinguirlas para ofrecer una propuesta de intervención adecuada (Nation & Snowling, 1997: O'Reilly et al., 2014). Keenan et al. (2008) ponen al descubierto la escasa correlación que existe entre test que supuestamente miden la comprensión lectora, dado a que probablemente miden aspectos diferentes bajo el multifacético constructo de la comprensión. Afortunadamente, algunos nuevos test superan estas limitaciones, como el Test de Evaluación de la Comprensión (TEC), cuyos destinatarios son niños de 10 a 16 años (Vidal-Abarca et al., 2007; ver Martínez et al., 2009, para una versión electrónica). Otro reciente test es Evaluación para la Comprensión Lectora (ECOMPLE-SE) para estudiantes de secundaria (Olmos Albacete et al., 2016), que, sin derivar de una teoría de la comprensión, cuenta con un modelo de evaluación. En Chile, entre las pruebas más usadas están las Pruebas de Comprensión Lectora y Producción de Textos (CL-PT) de kínder a 4º básico (Medina et al., 2011), de 5º a 8º básico (Fundación Educacional Arauco et al., 2010) y la Prueba de Comprensión Lectora de Complejidad Lingüística Progresiva (CLP; Alliende et al., 2004) de 1º a 8º básico, pero no se basan en teorías explícitas y aceptadas. A nivel prescolar empiezan a crearse en Chile pruebas sólidas, aunque todavía no estandarizadas ni comercializadas (Strasser et al., 2010).

El test que se presenta aquí está dirigido a niños de 3º a 6º año de básica. Incluye a niños de 3º de básica, porque es cuando cabe esperar un nivel de fluidez lectora suficiente para evaluar la comprensión lectora (ya suelen haber aprendido a leer y leen para aprender) y cuando aumenta tanto la brecha entre malos y buenos lectores, especialmente en estudiantes de bajo nivel socioeconómico (NSE; Kieffer, 2012), como la brecha en el rendimiento académico en general (Sirin, 2005). Por otro lado, el CLIP llega a 6º año de básica, porque es todavía una edad propicia para la enseñanza específica de la comprensión lectora y prepara para las exigencias de los textos de cursos superiores.

El objetivo de este estudio fue analizar las propiedades psicométricas del CLIP en una muestra de estudiantes chilenos de educación básica. Este objetivo incluye un análisis de la consistencia interna de las puntuaciones, de la estructura factorial y del funcionamiento diferencial de los ítems (DIF por sus siglas en inglés). Se esperaba encontrar una estructura factorial acorde con la naturaleza de las demandas de procesamiento que exige el CLIP. Como constructores de un test educativo, una preocupación es lograr equidad en las mediciones educativas donde participan grupos minoritarios. Esta preocupación recae también en los test de comprensión lectora, porque el contenido de los textos puede jugar un papel esencial en cómo se comprenden (Gafni, 1990). En este trabajo se examinaron dos fuentes de DIF: el género y el NSE. Se esperaba no encontrar ítems que funcionaran diferencialmente en grupos del mismo nivel de rendimiento, pero diferente género o NSE.

Método

Participantes

Este estudio fue realizado como parte de un proyecto de colaboración entre universidades chilenas y españolas para realizar proyectos y programas de mejora de la comprensión lectora basados en la evidencia científica y destinados a las escuelas de las regiones con más necesidades de apoyo (Fuentes Monsálves, 2009). Una de esas regiones en Chile es la Araucanía, donde secularmente se han observado los peores resultados en lectura del país, según la Agencia de Calidad de la Educación (2018). En este marco, se invitó a participar a colegios de las comunas de Villarrica y Loncoche. La muestra fue por conveniencia, compuesta por estudiantes de aquellos centros que se prestaron a colaborar.

Inicialmente participaron 2277 estudiantes, de los que 319 (14%) fueron excluidos por errores u omisiones en sus respuestas, por problemas significativos, corroborados por los administradores (e.g., estudiantes con necesidades educativas especiales) o por falta de consentimiento informado. La muestra final estuvo formada por 1958 estudiantes de 10 colegios de esas dos comunas (7 particulares subvencionados y 3 municipales) de niños entre 3º y 6º básico sin problemas sensoriales o cognitivos importantes. De hecho, ningún niño fue considerado con necesidades educativas especiales, según informaron los profesores a los coordinadores de los colegios que participaron en la recogida de datos del CLIP. Se utilizaron los cuestionarios y pruebas empleados en los colegios chilenos para determinar el NSE y el rendimiento académico y que se resumen en el llamado Índice de Vulnerabilidad Educativa (IVE) que indica el NSE del estudiante y el SIMCE, que indica el logro de los objetivos en lenguaje y matemática del centro. En la Tabla 1 se muestra que los participantes de este estudio se clasifican con un IVE medio y bajo y estudian en centros que alcanzaron un SIMCE intermedio y avanzado.

Diseño y Características del CLIP

La versión del CLIP que se examina aquí surge tras cuatro versiones anteriores que fueron administradas a muestras españolas entre los años 2003 y 2010 y de las que se seleccionaron los ítems que mejores índices de discriminación mostraron. Se explican a continuación las decisiones tomadas para construir el CLIP.

En cuanto a los criterios de elaboración de los textos, en primer lugar, se optó por crear textos que permitieran manipular las variables que influyen en la comprensión lectora. Entre otras, hubo una modificación general que consistió en eliminar elementos que facilitan la comprensión. Así, por un lado, se redujo el uso de señales estructurales (e.g., *el problema es*; Meyer, 1985), que pueden funcionar como señales facilitadoras de la comprensión (Goldman & Rakestraw Jr., 2000). Por otro lado, se cuidó que los textos tuvieran las ideas principales implícitas, evitando el uso de frases temáticas y títulos, y que estas ideas no estuvieran colocadas en la primera frase, como recomiendan algunos estudios (e.g., Surber, 2001), ya que las primeras frases de un párrafo a menudo son interpretadas como el tema o la idea principal (Goldman & Rakestraw Jr., 2000). Todas estas medidas referidas a la calidad, en conjunto, fueron dirigidas a obligar al lector a hacer una revisión del texto entero para dar con lo importante y así permitir evaluar si el lector integra las ideas del texto, una de las acciones que más cuesta a los malos comprendedores (Long & Chong, 2001) y que, por tanto, más puede ayudar a diferenciarlos de los buenos comprendedores.

Tabla 1
Edad Media (y Desviación Estándar) y Número de Niños por Sexo, Edad, IVE y SIMCE, según Curso

Curso	Sexo		Edad	IVE		SIMCE	
	V	M	*M* (*DE*)	Medio	Bajo	Intermedio	Avanzado
3º (*n* = 497)	245	252	8,05 (0,31)	434	63	354	143
4º (*n* = 503)	274	229	9,06 (0,35)	438	65	348	155
5º (*n* = 461)	227	234	10,11 (0,44)	398	63	327	134
6º (*n* = 497)	223	274	11,09 (0,39)	431	66	355	142
Total (*n* = 1958)	969	989	9,57 (1,20)	1701	257	1384	574

En segundo lugar, se optó por construir textos expositivos y narrativos, categoría que ha recibido apoyo teórico y empírico. Los textos narrativos son generalmente fáciles, están más próximos al discurso oral, su contenido es más familiar, personal o concreto (e.g., colección de episodios con protagonistas y que tienen metas e intenciones), inducen experiencia en el lector, divierten y entretienen y su estructura es sencilla. Los textos expositivos, en cambio, son generalmente difíciles, propios de la educación formal, su contenido es poco familiar, impersonal o abstracto (e.g., ideas, argumentos e instrucciones), inducen conocimiento en el lector, su lectura va acompañada de cierta sensación de esfuerzo y su estructura es compleja (de Vega et al., 1999). Otras diferencias radican en su propósito y características (Duke, 2009; Palincsar & Duke, 2004). Así, los textos narrativos pretenden contar una historia verdadera y se caracterizan por transmitir eventos en orden cronológico, presentar problemas y soluciones y utilizar dispositivos como fotografías o artefactos de un evento, mientras que los textos expositivos pretenden transmitir información sobre el mundo natural o social y se caracterizan por usar patrones organizacionales específicos, como comparar/contrastar, incluir definiciones o explicaciones de palabras que pueden ser desconocidas y emplear gráficos, como diagramas para transmitir información.

Los textos narrativos se construyeron teniendo en cuenta las partes descritas por van Dijk (1978/1992), principalmente, el suceso, formado por la complicación, que refleja la secuencia de acciones, normalmente a lo largo de más de una oración, y la resolución, que puede expresar el éxito o fracaso de las acciones realizadas. Otras partes son el marco, el episodio, la trama, la evaluación, la moraleja —aunque esta puede no estar en todos los textos— y la historia. Teniendo esto en cuenta, se construyeron algunos textos siguiendo el orden esperado, complicación-resolución, y otros alterando este orden, resolución-complicación.

Los textos expositivos se construyeron usando la clasificación de Kobayashi (2002), quien diferencia cuatro tipos de estructura: problema-solución, causa-efecto, descripción y asociación, ordenados de más a menos coherentes, donde las ideas están de más integradas a menos integradas. Si bien estas estructuras han sido un criterio de gran utilidad para graduar la complejidad de los textos expositivos, no se quiere obviar la complejidad que en él radica, pues las estructuras nunca aparecen de forma pura y a veces ni siquiera se alcanza un consenso total sobre qué tipo de estructura tiene un texto determinado.

Por último, se introdujeron textos cortos (entre 150 y 250 palabras) y textos largos (entre 400 y 500 palabras). Esta diversidad permitió conciliar varias necesidades. Por una parte, la presencia de textos cortos, que pueden ser leídos con relativa brevedad, posibilitó la evaluación de textos de variadas estructuras y temáticas. Por otra parte, la introducción de textos largos posibilitó la evaluación de textos más naturales o ajustados a la realidad de las aulas. Además, los textos largos parecen medir la comprensión lectora con mayor independencia de la fluidez lectora que los textos cortos (Keenan et al., 2008). Los textos, en origen, fueron creados en España, por lo que posteriormente fueron analizados por profesores de primaria y universitarios de Chile que actuaron como jueces. En dicho análisis, además de los criterios de calidad, interés y adecuación a la edad, se tuvo en cuenta que fueran textos adecuados a las distintas culturas. Así, se realizó una adaptación del léxico a la idiosincrasia chilena, en base a las indicaciones suministradas por dos profesores chilenos con más de 10 años de experiencia docente (ver Tabla 2 para más detalle de los textos).

En cuanto a los criterios de elaboración de las preguntas, si bien el modelo de construcción-integración especifica los niveles de representación que un lector puede alcanzar y los procesos cognitivos que requiere, en la práctica qué procesos medir y cómo medirlos es complejo. Así, por ejemplo, diferentes test se proponen medir procesos similares, pero con distintos nombres. En el TEC (Vidal-Abarca et al., 2007) miden cuatro procesos: captación de ideas explícitas, inferencias anafóricas, inferencias basadas en el conocimiento y formación de macroideas. En PISA (Schleicher & Tamassia, 2000) evalúan tres procesos: recuperación (requieren seleccionar y extraer información específica de los textos), interpretación (requieren conectar distintas informaciones a través de inferencias, siendo la fuente el propio texto) y reflexión-evaluación (requieren reflexionar más allá del texto). En PIRLS (Instituto Nacional de Evaluación Educativa, 2012) miden cuatro procesos: localización y obtención de información explícita; realización de inferencias directas; interpretación e integración de ideas e informaciones; y análisis y evaluación del contenido, el lenguaje y los elementos textuales. En SIMCE 2013 (Agencia de Calidad de la Educación, 2013) evalúan tres habilidades: localización de información, relación e interpretación de información y reflexión sobre el texto.

Tabla 2
Información de los Textos y Preguntas del CLIP

Texto	Tipo de género y estructura	Nº palabras	Escala Flesch*	Tipo de preguntas
Parte I de textos cortos				
Ramón	N. Orden directo	222	36	2L 2G 2I
Dolores de espalda	E. Problema-solución	238	26	2L 2G 2I
Pilar	N. Orden directo	167	49	2L 2G 2I
Tipos de delfines	E. Descriptivo	167	33	2L 2G 2I
Mario	N. Orden inverso	193	38	2L 2G 2I
Movimientos tierra	E. Asociativo	184	39	2L 2G 2I
Parte II de textos largos para 3º-4º				
Galeón pirata	N. Orden directo	378	34	5L 5G 5I
El pulpo	E. Problema-solución	427	31	5L 5G 5I
Parte II de textos largos para 5º-6º				
Las naranjas	N. Orden directo	527	45	5L 5G 5I
Las tortugas	E. Problema-solución	503	20	5L 5G 5I

Nota. N = Narrativo, E = Expositivo, L = Locales, G = Globales, I = Inferenciales.
*Prueba de nivel de facilidad de lectura de Flesch: 0 = Muy difícil, 100 = Muy fácil.

Por otro lado, respecto a cómo medir los diferentes niveles de representación, tampoco hay acuerdo. Algunos autores coinciden en que verificación de oraciones, recuerdo, parafraseo y resumen son tareas apropiadas para medir el texto base, porque se relacionan con evaluar información que está directamente dada en el texto (Kintsch, 1998). Sin embargo, estas mismas tareas también pueden ser buenos indicadores de modelo situacional, cuando exigen ir más allá del texto leído.

Son importantes también los resultados de la investigación sobre reconocidos test en lengua inglesa, en los que se muestra que algunos pueden responderse sin siquiera leer sus textos (Keenan et al., 2008) y que las habilidades de descifrado explican gran parte de la variabilidad en el rendimiento en el test, especialmente en 3º y 4º de primaria (Keenan & Meenan, 2014). Estos hallazgos son valiosos para sugerir tipos de preguntas y exigencias de los textos.

El CLIP incluye tres tipos de preguntas: locales, globales e inferenciales, siguiendo la clasificación de Magliano et al. (2007), que se dirigen a medir aspectos del texto base y del modelo situacional. El nivel de representación superficial no se mide, pues es un nivel muy básico, lejano al nivel de representación elaborado, que es propiamente la comprensión. Se eligió esta clasificación de preguntas, porque clarifica los procesos cognitivos subyacentes y porque puede ser útil para la enseñanza.

Las preguntas locales exigen al lector buscar y localizar información explícita en una frase o dos frases adyacentes y verificar cuál es la que mejor responde. Aquí se pueden encontrar las inferencias anafóricas o puente necesarias para mantener la cohesión a nivel local. Estas preguntas requieren pocas o nulas inferencias. Se asocia con el texto base. Se cuidó en hacer preguntas que usaran parafraseo y frases con abundante información para hacerlas exigentes.

Las preguntas globales exigen al lector mantener la coherencia a nivel global, en definitiva, captar lo importante y llegar al significado, pudiendo ser a nivel de frase, párrafo, grupo de párrafos o texto completo. En ocasiones, el texto facilita la construcción de esas inferencias (macrorregla de selección), pero a menudo no, con lo que el lector ha de hacer uso de su conocimiento previo, de otras inferencias (macrorreglas de generalización o integración) o de su conocimiento sobre las estructuras textuales. Estas preguntas se asocian con el texto base, pero también con el modelo de situación. En términos de procesos, el lector ha de buscar y localizar los segmentos apropiados del texto, construir un resumen del mismo y después seleccionar la opción que más se adecúa a ese resumen.

Las preguntas inferenciales exigen al lector ir más allá del texto con la información que ha extraído y con su conocimiento previo, junto con la capacidad de generar analogías, predecir y explicar. Este producto se

asocia con el modelo situacional. Entre este tipo de preguntas, la investigación diferencia dos clases, unas que son realizadas por los lectores expertos en el curso normal de lectura; otras que no son necesariamente construidas en el proceso de comprensión, porque realmente exigen al lector ir más allá del texto y le preguntan por contenidos que ni siquiera son mencionados en el texto. Este segundo tipo es el que se ha tratado de crear. Como señala la investigación, son las más difíciles no solo porque requieren al lector ir más allá de la información explícita, sino porque le exigen determinar en qué medida el texto soporta la respuesta alternativa similar a la inferencia correcta.

Con todo ello, la prueba se organizó en un cuadernillo con dos partes. La parte I estuvo compuesta por seis textos cortos (1º narrativo, 2º expositivo, 3º narrativo, 4º expositivo, 5º narrativo, 6º expositivo), iguales para todos los cursos, es decir, 3º, 4º, 5º y 6º. Cada texto tenía seis preguntas: dos locales (e.g., "Según el texto, los alumnos se levantaron corriendo porque…"), dos globales (e.g., "Elige la frase que mejor resume el contenido del texto") y dos inferenciales (e.g., "Teniendo en cuenta los sucesos de esta historia, ¿qué le recomendarías a Ramón?"). En total son 36 preguntas. La parte II estuvo compuesta por dos textos largos (1º narrativo, 2º expositivo), diferentes para 3º y 4º y para 5º y 6º. Cada texto tenía 15 preguntas: cinco locales, cinco globales y cinco inferenciales. En total son 30 preguntas. El test completo se puede obtener en http://villarrica.uc.cl/test. Los textos fueron ordenados de más fácil a más difícil, en base a los criterios mencionados en la elaboración, aunque los resultados obtenidos con la escala de Flesch no permiten apoyar esta graduación (ver Tabla 2).

Cada estudiante leyó en total ocho textos. El rango de puntuación total de la prueba fue de 0 a 66, otorgando 1 punto por cada ítem correcto.

Procedimiento

La recogida de datos tuvo lugar en el mes de abril de 2012, por universitarios entrenados. El primer paso fue solicitar a los padres o tutores de los estudiantes que constituían la muestra inicial la firma del consentimiento informado. Una vez seleccionados como participantes, los estudiantes respondieron al test en sus aulas y en grupo. No se limitó el tiempo de lectura, oscilando entre 60 y 80 minutos en 3º y 4º básico y entre 50 y 60 minutos en 5º y 6º básico, sin interrupción ni descansos, y podían consultar los textos, pues se ha mostrado que no hay diferencias entre responder con o sin consultar el texto; además, las diferencias entre buenos y malos lectores se mantienen constantes (Cain & Oakhill, 1999).

Análisis de Datos

En primer lugar, se realizaron análisis descriptivos de los resultados, análisis de la consistencia interna de las puntuaciones, nivel de la dificultad y poder discriminativo de los ítems.

Para el análisis de la confiabilidad de las puntuaciones se utilizó el coeficiente de consistencia interna Kuder-Richarson-20 para ítems dicotómicos. Para propósitos educativos y de investigación, una regla de oro es que la consistencia interna sea al menos 0,70 (Fraenkel et al., 2006). Para propósitos solo educativos cabe incluso aceptar un índice de consistencia interna de 0,60, siempre que en el test no recaigan decisiones graves, como la ubicación en un programa de educación especial. En cuanto al análisis de ítems, se examinó tanto el nivel de dificultad o proporción de aciertos como el poder discriminativo o correlación punto-biserial de cada ítem con el total del test corregido.

La evidencia de la validez de contenido o la relación entre el contenido del test y el constructo que se mide (Standards for educational and psychological testing, 2014) fue obtenida de dos formas en el CLIP. Primero, los ítems se crearon basados en el conocimiento actual sobre comprensión lectora y en una revisión de la mayoría de los test que miden el mismo constructo (comprensión lectora) en español y en inglés. Segundo, dos jueces expertos en la evaluación de la comprensión lectora valoraron aspectos del contenido de cada texto y, especialmente, sus respectivas preguntas o ítems, respondiendo a dos cuestiones sobre los ítems: qué nivel de procesamiento exigía cada pregunta (e.g., "superficial") y qué habilidad exigía (e.g., "exige localizar información al final del texto"). Para realizar esta tarea se ofreció a los jueces un cuestionario cualitativo, elaborado por los propios autores (disponible a quien lo solicite). Se empleó Kappa de Cohen para calcular la confiabilidad interjueces y la escala reportada por Viera y Garrett (2005) para interpretar el valor de este coeficiente.

La evidencia de la validez referida al constructo llevó a estudiar la estructura factorial del test a través de los dos tipos de análisis factorial: exploratorio (AFE) y confirmatorio (AFC). El uso del AFE es

recomendable, incluso contando con la existencia de una teoría previa, porque los resultados pueden indicar una estructura factorial diferente a la conceptualizada y explorada en los AFC. Contar con ambas fuentes de información es deseable en test nuevos y con apenas evidencia acumulada sobre su validez.

Así, la estructura interna del CLIP se examinó a través del AFE de los 66 ítems del test sin forzar una solución determinada. Para decidir el número de factores que debían ser extraídos se utilizó el método del análisis paralelo (Timmerman & Lorenzo-Seva, 2011). Se emplearon dos métodos de extracción, el de componentes principales y el de mínimos cuadrados, así como dos métodos de rotación de factores, Varimax y Promin.

Posteriormente, se realizó un AFC de las puntuaciones, con el fin de poner a prueba los tres tipos de preguntas correspondientes a los tres niveles de representación del texto. Partiendo de la matriz de covarianzas de las puntuaciones del test, se utilizó el procedimiento de la máxima verosimilitud. Siguiendo las orientaciones de expertos (Byrne, 2006; Hu & Bentler, 1999), fueron seleccionados como medidas de la bondad del ajuste de los modelos a los datos el estadístico χ^2, el índice comparativo de Bentler-Bonett (CFI) y la raíz del error cuadrático medio de aproximación (RMSEA). Valores de 0,90 o superiores para el CFI y de 0,06 o inferiores para el RMSEA son generalmente aceptados como un indicador de buen ajuste del modelo.

Por último, se examinó el DIF del CLIP para identificar aquellos con una probabilidad diferente de recibir respuestas correctas en grupos igualados en habilidad, pero diferentes en dos fuentes de posible sesgo: el género y el NSE (tal como es medido por el índice de vulnerabilidad). Para este análisis se utilizó el método de detección del DIF con la prueba Mantel-Hanszel (MH) adaptado en 1988 por Holland y Thayer (citados en Dorans & Holland, 1993).

Para los análisis estadísticos se utilizaron SPSS Statistics 26.0, JAMOVI (Jamovi Project, 2018), FACTOR versión 10.9.02 (Lorenzo-Seva & Ferrando, 2006; véase también Ferrando & Lorenzo-Seva, 2017) para el AFE, IBM SPSS AMOS versión 26.0 (Arbuckle, 2019) para el AFC y TiaPlus (Heuvelmans, 2003) para el análisis del DIF.

Resultados

Estadísticos Descriptivos, Consistencia Interna de las Puntuaciones y Análisis de Ítems

En la Tabla 3 se resume el rendimiento de los estudiantes por curso: media, desviación estándar y consistencia interna de las puntuaciones. La consistencia interna de las puntuaciones del CLIP es igual o mayor a 0,79 en todos los cursos y en el conjunto de la muestra, satisfaciendo, así, el criterio psicométrico de precisión en la medida.

El análisis de ítems reveló índices de dificultad entre 0,15 y 0,91, oscilando la mayoría de los ítems en el rango 0,30-0,70 (51 de los 66 ítems), siendo unos pocos los que mostraron valores extremos superiores a 0,91 o inferiores a 0,15. Los índices de discriminación, por su parte, oscilan en torno a valores de 0,20-0,40, habituales y satisfactorios en este tipo de test. Estos datos muestran que no cabe esperar grandes distorsiones en la determinación del número de factores comunes del test.

Tabla 3
Medias (Desviaciones Estándar) y Consistencia Interna por Curso

Curso	*M* (*DE*)	*r*
3º	22,43 (9,04)	0,84
4º	28,92 (9,22)	0,82
5º	28,69 (8,08)	0,79
6º	31,20 (8,78)	0,82
Total	28,80 (8,78)	0,84

Nota. *r* = Kuder-Richardson 20

Evidencias de la Validez Basadas en el Contenido

El análisis de las respuestas de los jueces sobre el ajuste de las preguntas del CLIP en términos de procesos y de habilidad reveló un coeficiente de 0,76, es decir, un substancial acuerdo respecto al nivel de procesamiento exigido por las preguntas, y un coeficiente de 0,62 o moderado acuerdo para las habilidades medidas en cada pregunta.

Evidencias de la Validez Basadas en la Estructura Interna

El test de Kaiser-Meyer-Olkin de adecuación de los datos para el análisis factorial exploratorio fue 0,89 y el test de esfericidad de Bartlett fue significativo, $\chi^2(2142, n = 1958) = 14644{,}0$, $p < 0{,}001$, indicando que la muestra y la matriz de correlaciones eran apropiadas para el análisis (Dziuban & Shirkey, 1974). Se creó una matriz de correlaciones tetracóricas de los ítems y se decidió someter todos los ítems a análisis. Los resultados del análisis paralelo indicaron que el número de factores a retener eran tres, considerando el percentil 95 de la varianza. La inspección del gráfico de sedimentación también sugería una solución de tres factores. Así, el análisis se limitó a tres factores (31,32% de la varianza total).

Se juzgó, en base a los factores, las cargas factoriales y lo interpretablead de los datos, que la extracción con el método de los mínimos cuadrados no ponderados (véase Timmerman & Lorenzo-Seva, 2011) y la rotación Promin eran la mejor solución. Los factores y sus respectivos ítems cuando la carga factorial fue mayor de 0,30 se presentan en la Tabla 4.

Hubo 24 ítems en el Factor I, 12 en el Factor II y 13 en el Factor III. El Factor I podría denominarse Comprensión de Textos Cortos, que exigen al lector generar procesos de comprensión posiblemente influidos por la calidad del descifrado (Keenan et al., 2008; Keenan & Meenan, 2014) y no le requieren de un procesamiento complejo. En contraste, el Factor II podría denominarse Comprensión de Textos Largos, que exigen procesos de comprensión quizá menos determinados por el descifrado, al ser más redundantes y dar ocasión al lector de encontrarse repetidamente con las palabras del texto. El Factor III podría denominarse Inferencias Basadas en el Conocimiento, pues recogió un grupo de ítems que exigían inferir respuestas basadas en conocimiento previo (e.g., "Probablemente, según el texto, ¿el pulpo se alimentará de…?").

A continuación, se realizó el AFC para evaluar la estructura subyacente al CLIP. Se evaluaron dos modelos del CLIP. El primer modelo fue el teórico e incluyó tres factores de primer orden interrelacionados: Procesamiento Local, Procesamiento Global y Procesamiento Inferencial, en correspondencia con los tres tipos de procesos presumiblemente evaluados por las preguntas del test. Teniendo en cuenta los resultados del AFE, el modelo teórico se puso a prueba en los textos cortos y en los largos, por separado. El segundo modelo fue un modelo unidimensional con un único factor, que se denominó Comprensión General.

Los índices de ajuste obtenidos para la presente muestra se presentan en la Tabla 5. El modelo teórico trifactorial demostró, en el caso de los textos cortos, un ajuste suficiente a los datos, con excepción de χ^2. Sin embargo, en los textos largos, el ajuste del modelo a los datos fue insatisfactorio, pues no satisfizo ninguno de los criterios estadísticos establecidos. El modelo unidimensional demostró un ajuste semejante al del modelo trifactorial, aunque, en el caso de los textos largos, superó al modelo trifactorial en cuanto al ajuste a los datos, pues al menos uno de los tres criterios establecidos fue satisfecho.

Evidencias Basadas en las Consecuencias de la Evaluación

En el análisis del DIF, cuando la fuente de sesgo fue el género, los varones fueron el grupo focal y las mujeres el grupo referencial, en base al rendimiento superior de las mujeres en test de comprensión lectora registrado en las evaluaciones internacionales (OECD, 2015). En el análisis del NSE como fuente de sesgo, el grupo focal fue el grupo con un IVE medio (es decir, mayor vulnerabilidad) y el referencial el de índice bajo (menor vulnerabilidad). Puesto que el criterio de referencia en este método de detección del DIF es la puntuación total en el test, se realizó, por separado, para cada fuente de sesgo, un primer examen del DIF para determinar si era necesario refinar el criterio, eliminando ítems con un DIF considerable. Este paso de refinamiento del criterio no fue necesario pues ningún ítem presentó DIF en este examen inicial.

Los resultados del análisis del DIF se presentan en la Tabla 6, donde se observa que no hubo ningún ítem cuya z alcanzara el valor crítico de 1,96, lo que indica que no hay diferencias entre los grupos comparados por género ni por NSE, cuando fueron igualados por nivel de habilidad (el programa empleado en este análisis divide la muestra total en cuatro niveles de habilidad).

Tabla 4

Matriz de Cargas Factoriales de los Ítems en la Extracción de Mínimos Cuadrados No Ponderados y Rotación Promin

Ítem	Tipo de texto y pregunta	Factor 1	Factor 2	Factor 3
1	cnl	0,439		
2	cnl	0,495		
3	cng	0,420		
4	cng			
5	cni	0,353		
6	cni			
7	cel			0,429
9	ceg	0,426		
10	ceg	0,432		
11	cei	0,345		
12	cei	0,382		
13	cnl	0,582		
14	cnl	0,405		
15	cng			0,302
17	cni	0,472		
18	cni	0,448		
19	cel	0,548		
20	cel	0,492		
21	ceg	0,486		
22	ceg			0,460
24	cei			0,343
25	cnl	0,470		
26	cnl	0,342		
28	cng			0,364
30	cni			0,417
31	cel	0,649		
32	cel	0,633		
33	ceg	0,371		
34	ceg	0,584		
35	cei	0,456		
36	lnl		0,491	
37	lnl		0,508	
38	lnl	0,326		
39	lnl		0,303	
44	lng		0,534	
45	lng			0,521
46	lni		0,693	
47	lni			0,375
48	lni			0,549
49	lni		0,441	
50	lni		0,772	
51	lel		0,580	
53	lel		0,487	
54	lel	0,302		
56	leg		0,548	
57	leg		0,319	
58	leg		0,318	
61	lei			0,673
62	lei			0,521
63	lei			0,551
64	lei		0,372	

Nota. En tipo de texto y pregunta, la primera letra indica texto corto (c) o largo (l), la segunda, narrativo (n) o expositivo (e) y la tercera, el tipo de pregunta, local (l), global (g) o inferencial (i).

Tabla 5
Bondad del Ajuste de los Modelos Basados en la Teoría

Modelo	χ^2	*gl*	*p*	CFI	RMSEA
Unidimensional					
Textos cortos	1256	594	< 0,001	0,895	0,024
Textos largos	2096	405	< 0,001	0,663	0,044
Tres factores correlacionados					
Textos cortos	1251	591	< 0,001	0,895	0,024
Textos largos	5182	404	< 0,001	0,048	0,078

Tabla 6
Resultados del Análisis del DIF Según Sexo y NSE

Ítem	Sexo		Índice de vulnerabilidad	
	Mantel-Haenszel	*z*	Mantel-Haenszel	*z*
1	0,7920	-0,6909	0,9542	-0,0892
2	1,2563	0,9657	0,8171	-0,5795
3	1,1828	0,7364	0,8834	-0,3663
4	0,9859	-0,0669	0,9766	-0,0749
5	0,8091	-0,9729	0,9217	-0,2545
6	1,1376	0,5706	0,9975	-0,0076
7	0,9374	-0,2741	1,6162	1,3724
8	0,9594	-0,1718	0,6589	-1,1029
9	0,8183	-0,8717	1,1134	0,3026
10	0,7984	-0,8720	0,7673	-0,6782
11	0,9137	-0,3319	0,8750	-0,3298
12	1,1210	0,4534	0,7565	-0,7477
13	0,9158	-0,3790	1,3767	0,9359
14	0,7513	-1,1935	1,0649	0,1668
15	0,9247	-0,3175	1,1041	0,2830
16	1,0353	0,1490	1,0070	0,0205
17	1,0361	0,1516	0,9737	-0,0784
18	0,7658	-1,1998	0,8986	-0,3256
19	0,9805	-0,0812	0,9334	-0,1842
20	0,8486	-0,6952	1,3707	0,9006
21	0,7677	-1,1606	1,0085	0,0254
22	1,2881	1,1180	1,4243	0,9865
23	0,9449	-0,1870	0,8767	-0,2813
24	1,2084	0,7582	1,0394	0,1071
25	0,7434	-1,1818	1,2753	0,6007
26	0,8493	-0,7553	0,7812	-0,7622
27	0,8630	-0,5555	1,2243	0,5395
28	1,2430	0,9045	1,1354	0,3694
29	1,0499	0,1812	0,7739	-0,6388
30	0,9093	-0,4149	1,1439	0,4047
31	0,8160	-0,8195	1,2799	0,7091
32	0,8582	-0,6644	1,8776	1,8590
33	0,9544	-0,1980	1,0679	0,1940
34	0,9447	-0,2343	1,2789	0,6874
35	0,7251	-1,4270	1,0251	0,0734

(continúa)

Tabla 6 (Conclusión)
Resultados del Análisis del DIF Según Sexo y NSE

Ítem	Sexo		Índice de vulnerabilidad	
	Mantel-Haenszel	*z*	Mantel-Haenszel	*z*
36	1,1090	0,4456	1,3152	0,8300
37	1,2221	0,9174	0,6786	-1,2018
38	0,7946	-1,0244	1,1428	0,3899
39	0,8958	-0,4940	0,7542	-0,8480
40	0,7529	-1,0882	1,0420	0,1013
41	0,7263	-1,2217	0,9203	-0,2038
42	0,9914	-0,0376	0,9193	-0,2469
43	0,8843	-0,5512	0,9241	-0,2384
44	0,8034	-0,9602	0,9305	-0,2089
45	0,8123	-0,9338	0,8623	-0,4547
46	1,0217	0,0959	1,2115	0,5753
47	1,2751	1,0461	0,9918	-0,0246
48	0,6776	-1,7314	1,2153	0,5662
49	0,8283	-0,7210	0,7055	-0,8757
50	1,3430	1,3297	0,6093	-1,4392
51	1,2210	0,8841	1,0672	0,1972
52	1,0826	0,3700	1,0368	0,1126
53	1,2045	0,8553	0,9580	-0,1341
54	1,2597	0,9913	0,7730	-0,7561
55	1,0875	0,3575	0,8637	-0,4225
56	1,1042	0,4402	1,0792	0,2256
57	1,3901	1,5475	0,8854	-0,3888
58	1,1794	0,7302	0,8801	-0,3848
59	0,9702	-0,1376	0,9571	-0,1351
60	1,1264	0,4556	0,8095	-0,5284
61	1,0675	0,2908	1,2901	0,7955
62	1,0966	0,4162	1,0962	0,2753
63	0,9548	-0,1875	0,9699	-0,0854
64	1,0448	0,1894	1,1441	0,4057
65	1,2267	0,9092	0,7441	-0,8802
66	0,9880	-0,0551	1,0664	0,1972

Discusión

El objetivo de este estudio fue examinar las propiedades psicométricas del test CLIP, recientemente creado para medir la comprensión lectora de textos expositivos y narrativos, en una muestra de estudiantes chilenos de 3º a 6º de educación básica de las comunas de Villarrica y Loncoche de la IX Región de Chile. Los resultados mostraron que las puntuaciones del CLIP en la muestra estudiada alcanzan una consistencia interna satisfactoria, con índices iguales o superiores a 0,79. El análisis de la dificultad de los ítems mostró que la mayoría de los ítems estuvo entre valores de 0,30 y 0,70, siendo unos pocos los que mostraron valores extremos superiores a 0,91 o inferiores a 0,15.

Los resultados del AFE mostraron una estructura interna de tres factores que se corresponden aproximadamente con la longitud de los textos y las preguntas inferenciales de conocimiento. Este hallazgo coincide parcialmente con el de Keenan y Meenan (2014) sobre las diferencias en el procesamiento de textos cortos y largos, debido al peso del descifrado en los procesos de comprensión, siendo los textos cortos más influidos por el reconocimiento visual de palabras que los textos largos. Por otra parte, los resultados del AFC son aparentemente decepcionantes, por no revelar un buen ajuste del modelo teórico propuesto, este es, un modelo trifactorial representado por tres factores correlacionados que se refieren a los tres tipos de procesamiento que exigen las preguntas del CLIP: local, global e inferencial. El modelo unidimensional con

el que se comparó el modelo teórico alcanzó un ajuste pobre, pero no muy diferente al modelo de tres factores. Este resultado coincide con estudios previos con otros test de comprensión lectora que se han sometido al AFC. Por ejemplo, Vidal-Abarca et al. (2007) encuentran como mejor modelo de ajuste a los datos un modelo unifactorial con un factor general que un modelo tetrafactorial que representaría a las cuatro formas de inferencia que incorpora el TEC. También Olmos Albacete et al. (2016) encuentran que un modelo de factor general se ajusta satisfactoriamente a los datos obtenidos en el AFC del ECOMPLE-SE. Coincidiendo con el argumento de estos mismos autores, es probable que, siendo la comprensión lectora un proceso cíclico, resulte difícil hallar factores de estructura independiente. Este resultado y los del AFE justifican, en parte, el cálculo de una puntuación total en el test. Ello no obsta para que el análisis de cada parte del test, incluso de cada ítem, sea una oportunidad para analizar y enseñar los procesos que subyacen a la comprensión de los textos. Pero la validez de este uso de los resultados del CLIP debe ponerse a prueba en futuros estudios de enseñanza de la comprensión.

Un resultado importante es que ningún ítem del CLIP muestra DIF en los participantes del estudio. No cabe generalizar este resultado, pero puede considerarse bastante robusto, gracias al tamaño de la muestra y la suficiente potencia estadística para detectar sesgos, en caso de haberlos. Cabía esperar la ausencia de sesgos en los ítems, debido a que el proceso de elaboración de los textos del CLIP estuvo guiado por las recomendaciones actuales al respecto (Standards for educational and psychological testing, 2014). Todos los contenidos del CLIP fueron cuidadosamente formulados y revisados para que no incluyeran, en lo posible, contenidos estereotipados o más característicos de ciertos grupos de la población. A pesar de este buen resultado, puesto que el criterio para examinar el sesgo de los ítems fue interno (i.e., la puntuación total en el test), en lugar de un criterio externo (e.g., la puntuación en otro test), hay que tener en cuenta que la única inferencia válida respecto a la ausencia de sesgos del CLIP es en el contexto del resto de ítems del test. El estudio DIF no logra escapar de una cierta circularidad que no exime de llevar a cabo un análisis racional, por parte de expertos, de los contenidos del test, tal como se llevó a cabo durante el análisis de contenido del CLIP con conclusiones satisfactorias.

Las limitaciones de este trabajo son varias. Una, el presente estudio fue realizado en una región chilena, lo que supone una obvia limitación a la generalización de los resultados. Es necesario seguir investigando las propiedades psicométricas del CLIP en muestras de estudiantes de otras regiones chilenas y de otros países de habla hispana. Otra limitación es que este trabajo solo ofrece una forma del test, cuando un test como el presente, destinado de evaluar repetidas veces a la misma población, exige crear formas paralelas, además de actualizar el contenido de los textos, con el fin de reducir el efecto de la práctica (Hausknecht et al., 2007). Otra limitación inherente a cualquier primer estudio es la ausencia de datos sobre otras facetas de la validez predictiva y de constructo. Es importante que futuros estudios sobre la validez de las puntuaciones del CLIP se centren en probar su validez como medida del efecto de la enseñanza de la comprensión lectora (O'Reilly et al., 2014; Scott, 2009), que es el fin para el que fue creado. Este es el desafío pendiente de la mayoría de las medidas de comprensión lectora.

Referencias

Afflerbach, P., Cho, B. -Y., Kim, J. -Y., Crassas, M. E. & Doyle, B. (2013). Reading: What else matters besides strategies and skills? *The Reading Teacher*, *66*(6), 440-448. https://doi.org/10.1002/TRTR.1146

Agencia de Calidad de la Educación. (2013). *Orientaciones para docentes educación básica SIMCE 2013*. Gobierno de Chile. http://archivos.agenciaeducacion.cl/biblioteca_digital_historica/orientacion/2013/orien_docbasica_2013.pdf

Agencia de Calidad de la Educación. (2018). *Informe de resultados Estudio Nacional: Lectura 2° básico 2017*. Gobierno de Chile. http://archivos.agenciaeducacion.cl/IRE_LECTURA_2018_2BASICA_WEB_ALTA_11_JUL.pdf

Alexander, P. A. (2003). Profiling the developing reader: The interplay of knowledge, interest, and strategic processing. En C. M. Fairbanks, J. Worthy, B. Maloch, J. V. Hoffman & D. L. Schallert (Eds.), *The fifty-first yearbook of the National Reading Conference* (pp. 47-65). National Reading Conference.

Alexander, P. A. & The Disciplined Reading and Learning Research Laboratory. (2012). Reading into the future: Competence for the 21st Century. *Educational Psychologist*, *47*(4), 259-280. https://doi.org/10.1080/00461520.2012.722511

Alliende, F., Condemarín, M. & Milicic, N. (2004). *Prueba CLP (Comprensión Lectora de Complejidad Lingüística Progresiva)*. Pontificia Universidad Católica de Chile.

Arbuckle, J. L. (2019). *IBM SPSS Amos 26*: *User's guide* (Versión 26.0) [Software computacional]. IBM. https://www.ibm.com/support/pages/spss-amos-26-documentation

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2ª ed.). Lawrence Erlbaum.

Cain, K. (2010). *Reading development and difficulties*. British Psychological Society & Blackwell.

Cain, K. & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*(5-6), 489-503. https://doi.org/10.1023/A:1008084120205

de Vega, M., Díaz, J. M. & León, I. (1999). Procesamiento del discurso. En M. de Vega y F. Cuetos (Coords.), *Psicolingüística del español* (pp. 271-305). Trotta.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. En P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Lawrence Erlbaum.

Duke, N. K. (2009). Informational text and young children: When, why, what, where, and how. *National Geographic Learning: Best Practices in Science Education*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.611.7588&rep=rep1&type=pdf

Dziuban, C. D. & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin*, *81*(6), 358-361. https://doi.org/10.1037/h0036316

Ferrando, P. J. & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*(2), 236-240. https://doi.org/10.7334/psicothema2016.304

Fraenkel, J. R., Wallen, N. E. & Hyun, H. H. (2006). *How to design and evaluate research in education* (6ª ed.). McGraw-Hill.

Fuentes Monsálves, L. I. (2009). Diagnóstico de comprensión lectora en educación básica en Villarrica y Loncoche, Chile. *Perfiles Educativos, 31*(125), 23-37. https://doi.org/10.22201/iisue.24486167e.2009.125.18844

Fundación Educacional Arauco, Medina, A. & Gajardo, A. M. (2010). *Pruebas de comprensión lectora y producción de textos (CL-PT): 5º a 8º año básico. Marco conceptual y manual de aplicación y corrección*. Ediciones UC.

Gafni, N. (25-27 abril de 1990). *Differential item functioning: Performance by sex on reading comprehension tests* [Presentación de ponencia]. 9th Annual Meeting of the Academic Committee for Research on Language Testing, Kiryat Anavim, Israel. https://files.eric.ed.gov/fulltext/ED331844.pdf

Goldman, S. R. & Rakestraw Jr., J. A. (2000). Structural aspects of constructing meaning from text. En M. L. Kamil, P. B. Mosenthal, P. D. Pearson & R. Barr (Eds.), *Handbook of reading research Volume III* (pp. 314-335). Lawrence Erlbaum.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T. & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373-385. https://doi.org/10.1037/0021-9010.92.2.373

Heuvelmans, T. (2003). *TiaPlus* [Software computacional]. CITO. https://www.cito.nl/kennis-en-innovatie/psychometrisch-onderzoek-en-dienstverlening/tools-voor-toetsontwikkelaars/tools-voor-toetsanalyse/tia-plus

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(4), 1-55. https://doi.org/10.1080/10705519909540118

Ingels, S. J., Burns, L. J., Chen, X., Cataldi, E. F. & Charleston, S. (2005). *A profile of the American high school sophomore in 2002: Initial results from the base year of the Education Longitudinal Study of 2002. Statistical analysis report* (NCES 2005-338). U.S. Department of Education, National Center for Education Statistics. https://nces.ed.gov/pubs2005/2005338.pdf

Instituto Nacional de Evaluación Educativa. (2012). *PIRLS-TIMSS 2011. Estudio internacional de progreso en comprensión lectora, matemáticas y ciencias. IEA. Volumen I: Informe español*. Gobierno de España, Ministerio de Educación, Cultura y Deporte, Secretaría de Estado de Educación, Formación Profesional y Universidades, Dirección General de Evaluación y Cooperación Territorial. https://sede.educacion.gob.es/publiventa/descarga.action?f_codigo_agc=15972

Jamovi Project (2018). *Jamovi* (Versión 0.9) [Software computacional]. https://www.jamovi.org

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Keenan, J. M., Betjemann, R. S. & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281-300. https://doi.org/10.1080/10888430802132279

Keenan, J. M. & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, *47*(2), 125-135. https://doi.org/10.1177/0022219412439326

Kieffer, M. J. (2012). Before and after third grade: Longitudinal evidence for the shifting role of socioeconomic status in reading growth. *Reading and Writing*, *25*(7), 1725-1746. https://doi.org/10.1007/s11145-011-9339-2

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kispal, A. (2008). *Effective teaching of inference skills for reading: Literature review* (Research Report DCSF-RR031). National Foundation for Educational Research, Department for Children, Schools and Family. https://dera.ioe.ac.uk/7918/1/DCSF-RR031.pdf

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, *19*(2), 193-220. https://doi.org/10.1191/0265532202lt227oa

Lauff, E., Ingels, S. J. & Christopher, E. (2015). *Education longitudinal study of 2002 (ELS:2002): A first look at the postsecondary transcripts of 2002 high school sophomores*. U.S. Department of Education, Institute of Education, National Center for Education Statistics. https://nces.ed.gov/pubs2015/2015034.pdf

Long, D. L. & Chong, J. L. (2001). Comprehension skill and global coherence: A paradoxical picture of poor comprehenders' abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1424-1429. https://doi.org/10.1037/0278-7393.27.6.1424

Lorenzo-Seva, U. & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods*, *38*(1), 88-91. https://doi.org/10.3758/bf03192753

Magliano, J. P., Millis, K., Ozuru, Y. & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. En D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107-136). Lawrence Erlbaum.

Martínez, T., Vidal-Abarca, E., Gil, L. & Gilabert, R. (2009). On-line assessment of comprehension processes. *The Spanish Journal of Psychology*, *12*(1), 308-319. https://doi.org/10.1017/S1138741600001700

Martínez, T., Vidal-Abarca, E., Sellés, P. & Gilabert, R. (2008). Evaluación de estrategias y procesos de comprensión: el Test de Procesos de Comprensión. *Infancia y Aprendizaje*, *31*(3), 319-332. https://doi.org/10.1174/021037008785702956

Medina, A., Gajardo, A. M. & Fundación Educacional Arauco. (2011). *Pruebas de comprensión lectora y producción de textos (CL-PT): Kinder a 4º básico. Marco conceptual y manual de aplicación y corrección*. Ediciones UC.

Meyer, B. J. F. (1985). Prose analysis: Purposes, procedures, and problems. En B. K. Britton & J. B. Black (Eds.), *Understanding expository text: A theoretical and practical handbook for analyzing explanatory text* (pp. 11-64). Lawrence Erlbaum.

Montanero Fernández, M. (2004). Cómo evaluar la comprensión lectora: alternativas y limitaciones. *Revista de Educación de España*, *335*, 415-427. http://www.educacionyfp.gob.es/dam/jcr:e43c09bf-2cc4-4315-8f0f-49012e97f470/re33526-pdf.pdf

Nation, K. & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*(3), 359-370. https://doi.org/10.1111/j.2044-8279.1997.tb01250.x

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Institute of Child Health and Human Development. https://www.nichd.nih.gov/publications/pubs/nrp/smallbook

Olmos Albacete, R., León Cascón, J. A., Martín Arnal, L. A., Moreno Pérez, J. D., Escudero Domínguez, I. & Sánchez Sánchez, F. (2016). Psychometric properties of the Reading Comprehension Test ECOMPLEC-SEC. *Psicothema*, *28*(1), 89-95. https:/doi.org/10.7334/psicothema2015.92

O'Reilly, T., Sabatini, J. & Wang, Z., (2018). Using scenario-based assessments to measure deep learning. En K. Millis, D. L. Long, J. P. Magliano & K. Weimer (Eds.), *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp. 197-208). Routledge. https://doi.org/10.4324/9781315109503-16

O'Reilly, T., Weeks, J., Sabatini, J., Halderman, L. & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review*, *26*(3), 403-424. https://doi.org/10.1007/s10648-014-9269-z

Organisation for Economic Co-operation and Development. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. https://doi.org/10.1787/9789264229945-en

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. https://doi.org/10.1787/9789264266490-en

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 results (Volume I): What students know and can do*. https://doi.org/10.1787/5f07c754-en

Palincsar, A. S. & Duke, N. K. (2004). The role of text and text-reader interactions in young children's reading development and achievement. *The Elementary School Journal*, *105*(2), 183-197. https://doi.org/10.1086/428864

Schleicher, A. & Tamassia, C. (2000). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy*. Organisation for Economic Co-operation and Development, Statistics and Indicators Division of the OECD Directorate for Education, Employment, Labour and Social Affairs . https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/33692793.pdf

Scott, S. E. (2009). *Knowledge for teaching reading comprehension: Mapping the terrain* (Tesis de Doctorado, University of Michigan). Deepblue. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/62201/sarascot_1.pdf?sequence=1&isAllowed=y

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417-453. https://doi.org/10.3102/00346543075003417

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. RAND Reading Study Group. U.S. Department of Education, Office of Educational Research and Improvement. http://www.rand.org/content/dam/rand/pubs/monograph_reports/2005/MR1465.pdf

Spencer, M. & Wagner, R. K. (2018). The comprehension problems of children with poor reading comprehension despite adequate decoding: A meta-analysis. *Review of Educational Research*, *88*(3), 366-400. https://doi.org/10.3102/0034654317749187

Standards for educational and psychological testing, American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). https://www.apa.org/science/programs/testing/standards

Strasser, K., Larraín, A., López de Lérida, S. & Lissi, M. R. (2010). La comprensión narrativa en edad preescolar: un instrumento para su medición. *Psykhe*, *19*(1), 75-87. https://doi.org/10.4067/S0718-22282010000100006

Surber, J. R. (2001). Effect of topic label repetition and importance on reading time and recall of text. *Journal of Educational Psychology*, *93*(2), 279-287. https://doi.org/10.1037/0022-0663.93.2.279

Timmerman, M. E. & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209-220. https://doi.org/10.1037/a0023353

van Dijk, T. A. (1992). *La ciencia del texto* (2ª reempresión; S. Hunzinger, Trad.). Paidós. (Obra original publicada en 1978)

Vidal-Abarca, E., Gilabert, R., Martínez, T., Sellés, P., Abad, N. & Ferrer, C. (2007). *TEC: Test de Estrategias de Comprensión*. Instituto Calasanz de Ciencias de la Educación.

Viera, A. J. & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine*, *37*(5), 360-363. http://www1.cs.columbia.edu/~julia/courses/CS6998/Interrater_agreement.Kappa_statistic.pdf

Fecha de recepción: Enero de 2018.
Fecha de aceptación: Noviembre 2019.