ISSN: 1989-0397

Revista Iberoamericana de Evaluación Educativa

Mayo 2017 - Volumen 10, número 1
https://doi.org/10.15366/riee2017.10.1

rinace.net/riee/
revistas.uam/riee

UNIVERSIDAD AUTÓNOMA DE MADRID

Red Iberoamericana de Investigación sobre Cambio y Eficacia Escolar
RINACE

CONSEJO EDITORIAL

ÍNDICE

Editorial

Sección Temática: Evaluación estandarizada

Temática libre

Editorial:

Evaluación Estandarizada

Standardized Assessment

Jesús Miguel Jornet Meliá *

Universidad de Valencia

La estandarización de la medida y evaluación en CC. Sociales en general y en CC. de la Educación en particular ha estado en el centro de los debates epistemológicos tanto respecto a los procesos de construcción del conocimiento como en los aplicados, por ejemplo, la evaluación.

Por estandarización entendemos el proceso de sistematización de todos los elementos de acercamiento a una acción de recogida e interpretación de información, de manera que se utilicen los mismos: instrumentos o técnicas, criterios de corrección y/o síntesis o análisis de la información y criterios de interpretación de la misma.

Estas estrategias se pueden aplicar tanto para clasificar instrumentos y técnicas de recogida de información, como para clasificar planes o programas de evaluación. No obstante, su mayor impacto, sin duda, lo tiene en el diseño de instrumentos y/o técnicas de recogida de información.

Siguiendo esta posición, y respecto a instrumentos y técnicas, podemos identificar tres grandes tipos de estrategias:

a) *Estandarización completa.* Implica que todos los elementos del instrumento de medida están sistematizados y se aplican de la misma manera para todas las personas. De este modo, se les presentan los mismos estímulos (reactivos o ítems), se les dan las mismas instrucciones de aplicación, se administran en un mismo tipo de situación (por ejemplo, si existe tiempo limitado o no), ellos tienen que ofrecer sus respuestas del mismo modo (por ejemplo, marcando o señalando en papel o seleccionando en ordenador, o escribiendo su respuesta), se corrigen de la misma manera y se interpretan las puntuaciones de acuerdo a los mismos criterios (estándares de interpretación, baremos, normas…).

b) *Estandarización parcial.* Generalmente tienen prefijados o sistematizados los modos de recogida de información, si bien en la forma de sintetizar o analizar la información existen posibilidades alternativas o mayor flexibilidad.

c) *Estandarización nula.* En este caso, ninguno de los elementos están prefijados y se recaba, sintetiza o analiza e interpreta la información de manera particular para cada caso.

La elección de uno u otro tipo de instrumento depende en muchas ocasiones de la posición epistemológica o metodológica del evaluador. No obstante, desde nuestro punto de vista, debería depender de las características del objeto de medición/evaluación y de

*Contacto: jornet@uv.es

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

la finalidad; de modo que nos situamos en una posición de complementariedad metodológica cuantitativa/cualitativa, reconociendo la complejidad de los hechos educativos y la necesidad de aunar esfuerzos desde diversas posturas para representar de manera adecuada la realidad objeto de estudio.

La estandarización ha estado defendida desde posiciones en las que se enfatiza la necesidad de plantear un mismo sistema de acercamiento a la realidad (sea en pruebas de rendimiento, escalas de actitudes o sistemas de evaluación de docentes, instituciones, programas…) con el fin de que las diferencias en relación a los resultados puedan imputarse a la diversidad del objeto o sujeto evaluado o a efectos debidos a una intervención sobre ellos, y no a posibles factores diferenciales derivados del modo en que han diseñado, aplicado o interpretado los instrumentos.

El supuesto de base es que debe mantenerse el conjunto de estímulos y/o modos de recogida de información de la misma manera con el fin de crear la misma situación. Las personas reaccionan ante la misma situación estimular de diversas maneras. Esa diversidad de respuesta caracteriza a las personas y no es debida a diferencias del instrumento o su uso. Este concepto, sin duda, es atractivo y parece lógico, pero la realidad nos demuestra que el hecho de estandarizar no siempre es adecuado y no asegura que se cumpla el supuesto mencionado, pues siempre existen factores que inciden de una u otra manera y que implican que cualquier medida contenga error, tal como señaló Spearman a comienzos del s. XX marcando el origen de la teoría clásica de construcción de tests.

Este planteamiento, sin embargo, ha sido característico de los procesos de recogida de información en los que, durante buena parte del siglo XX y el actual, se ha ido diferenciando entre instrumentos de medida, propiamente dichos, y las técnicas evaluativas. Los instrumentos de medida (tests psicológicos, pruebas de rendimiento, escalas de percepción y actitudes) están totalmente estandarizados, mientras que las técnicas evaluativas (cuestionarios, observación, entrevista…) pueden recorrer los tres niveles, si bien, lo más usual es que correspondan a procesos parcialmente estandarizados o con estandarización nula. En cualquier caso, podríamos imaginar un eje en el que en un extremo se situaran los instrumentos totalmente estandarizados -por ende supuestamente independientes del observador/evaluador-, y en el otro extremo, los no estandarizados –dependientes, en ese caso, de diversos factores, pero siendo central el rol del observador/evaluador-.

Todos los métodos y técnicas de recogida de información están sujetos a criterios de bondad: fiabilidad y validez. No obstante, por el tipo de metodología en que se sustenta cada procedimiento, la cuantificación de los criterios en indicadores que permitan informar de la fiabilidad de los mismos, o en evidencias que apoyen su validez, son más fácilmente identificables en los totalmente estandarizados. Ciertamente el control sobre los criterios de bondad de técnicas evaluativas resulta siempre más complejo, al no disponer de modelos de medida que los sustenten y por la necesidad de integrar informaciones más diversas, en las que el control sobre las mismas es más difícil de realizar. Ello no quiere decir que no posean tales características. Únicamente que resulta de mayor complejidad poder comprobarlas.

El problema de la validez, sin embargo, se identifica en todos los instrumentos y técnicas. Probablemente es el criterio más importante y a su vez más olvidado y que no se resuelve a través de modelos de medida. Recuérdese que la fiabilidad, en todo caso, es

una condición necesaria, pero nunca suficiente para la validez. Buena parte de los problemas que se imputan a las pruebas estandarizadas están relacionados precisamente con carencias en la validez. De ahí que objetivar sea una cuestión fundamental, que se enraíza en el sustrato de la validez y no queda nunca atendida por más que utilicemos los modelos de medición más avanzados. El problema es que, en muchas ocasiones, usuarios e incluso expertos creen que los modelos de medición aseguran todo y, por desgracia, no es así. En la base de muchos debates acerca de si la estandarización es una aproximación adecuada normalmente se identifican los problemas de fondo acerca de la validez.

Más allá de la evaluación de alumnado para procesos de selección o certificación, la estandarización también ha llegado a instaurarse en diversos procesos de evaluación, como por ejemplo la evaluación de docentes, instituciones y sistemas educativos. No obstante, si diseñar instrumentos de medida es complejo, establecer planes de evaluación (que deben integrar diversas informaciones, ser amplios y atender a diversas variables de entrada, proceso, contexto y resultado…) lo es más aún. Con todo, hay una tendencia clara en diversos tipos de planes –como por ejemplo, para evaluar instituciones o sistemas educativos– a reducir el concepto de calidad a la interpretación de los resultados del proceso educativo. En tales casos, es frecuente observar planes de evaluación que toman como eje vertebrador del análisis evaluativo el uso de pruebas estandarizadas de desempeño o logro educativos y, a lo sumo, integran otras variables recogidas a través de cuestionarios de contexto.

Sin pretender volver al debate sobre ello, pues nos hemos manifestado al respecto en diversos medios y ocasiones, es obvio que resultan planes muy parciales, en los que la validez suele estar en entredicho. Por ello, aunque la tendencia sea estandarizar también en este tipo de procesos, entendemos que más allá de la extracción de indicadores aislados de evaluación, la estandarización para situaciones más complejas (como son, por ejemplo, las instituciones o sistemas educativos, o la actuación docente…) es más que discutible. La dificultad de llevar el concepto de estandarización a programas o planes de evaluación radica, en definitiva, en las características del objeto de evaluación y en el concepto de calidad que quiera comprobarse. Por ello, a no ser que se reduzca a variables de resultado, es muy complicado intentar diseñar planes totalmente estandarizados. A lo sumo podrán integrarse algunos indicadores estandarizados en un plan que, por definición, deberá ser parcialmente estandarizado. Y, si se da lo anterior –reducir el plan a variables de resultado-, la validez de la evaluación será un factor deficitario, con toda seguridad.

En este número monográfico se integran diversos trabajos. Los tres primeros (de Agustín Tristán y Nancy Pedraza, Richard Phelps, y, finalmente, el de Manuel Fernández Navas, Noelia Alcaráz y Miguel Solá) presentan reflexiones de carácter teórico-metodológico que recogen diversas miradas de reflexión crítica acerca de la estandarización, desde posiciones más cercanas al análisis epistemológico hasta otras más impregnadas de ideología. En el siguiente artículo, Carolina Ruminot nos presenta un análisis sobre las consecuencias del SIMCE en Chile para el sistema educativo, basado en un análisis con docentes y proponiendo elementos derivados de ello para la intervención educativa. Pilar Rodríguez Morales nos muestra en su trabajo el diseño de una prueba de matemáticas y comprensión lectora para el ingreso de estudiantes de colectivos vulnerables en Uruguay, incluyendo aproximaciones de Teoría de Respuesta al Ítem. Una muestra de acercamiento de estandarización para evaluar la práctica docente, basada en observación y rúbricas y análisis sustentados sobre Teoría de la

Generalizabilidad, nos la aportan Luis H. Pedroza y Edna Luna, sin duda de gran interés para la valoración de las prácticas de aula. En los dos últimos trabajos, el de Cristina V. Hernández de la Toba y Joaquín Caso, y el de Carlos Sancho, Jesús M. Jornet y José González-Such, se presentan dos procesos de trabajo de estandarización sobre variables socio-afectivas. En el primer caso, sobre una escala de convivencia escolar y, en el segundo, sobre otra orientada a analizar el valor social que subjetivamente se da a la educación. En conjunto, se ofrece un recorrido sobre trabajos en torno a la estandarización que esperamos sean de interés. Agradecemos, desde estas líneas, a los autores por su trabajo, así como a la Revista Iberoamericana de Evaluación Educativa por abrir este tema a debate.

Sección Temática:

Evaluación Estandarizada

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 11-31.
https://doi.org/10.15366/riee2017.10.1.001

La Objetividad en las Pruebas Estandarizadas

Objectivity in Standardized Tests

Agustín Tristán López*
Nancy Yahibé Pedraza Corpus

Instituto de Evaluación e Ingeniería Avanzada (IEIA)

La objetividad es un atributo necesario que debe detallarse claramente para satisfacer los propósitos científicos de todo proyecto de evaluación en ciencias de la salud, ciencias sociales y educación, así como en cada una de las etapas de producción y uso de las pruebas estandarizadas. El valor de la objetividad para el desarrollo de las pruebas se refuerza al emplearse como herramienta de vigilancia que garantiza la neutralidad en los estímulos presentados. Se detallan cinco propiedades principales distintivas: especificidad, neutralidad, independencia, imparcialidad e impersonalidad, fundamentales para interpretar los resultados, eliminar o reducir los sesgos inducidos por la influencia de estereotipos y preferencias en el diseño del instrumento o en la apreciación de jueces, entre otros factores que pueden afectar el uso ético de los resultados de las pruebas. Se muestra que la objetividad es el primer atributo que debe definirse en una prueba estandarizada, distinguiendo las cualidades que le son propias para evitar asociarlas incorrectamente con la validez o la confiabilidad.

Palabras Claves: Objetividad, Pruebas estandarizadas, Validez, Confiabilidad.

Objectivity is a needed attribute of standardized tests in different areas, such as health, social sciences and education, and in each one of the phases of the development of a test, from its initial definition to the interpretation of outcomes. Objectivity ensures fairness of the test from its design up to the appraisal of the judges or evaluators and on the treatment of results, grounded on five main properties: specificity, neutrality, independence, impartiality and impersonality. Objectivity is fundamental for the interpretation of the outcomes, eliminating or reducing the presence of stereotypes and preferences that produce several types of bias that may affect the ethical use of the results of the test. Objectivity should be the first attribute to consider in a standardized test, as it improves the definition of the traits to evaluate permitting the distinction of characteristics that are mistakenly associated with validity and reliability.

Keywords: Objectivity, Standardized tests, Validity, Reliability.

*Contacto: atristan@ieia.com.mx

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 23 de octubre de 2016
1ª Evaluación: 15 de enero de 2017
Aceptado: 25 de febrero de 2017

1. Presentación

Las pruebas estandarizadas en educación siempre están en el ojo inquisitivo de funcionarios y autoridades públicas, de asociaciones de padres de familia, de docentes y de los estudiantes que deben resolver la prueba. No es propósito de este trabajo hacer la diatriba de estas pruebas ni tampoco su defensa, sino apuntar de manera breve las cualidades de la objetividad que, junto con la validez y la confiabilidad, configuran los tres atributos fundamentales para el diseño, administración, interpretación y uso de las pruebas estandarizadas.

En general los proyectos de evaluación se centran en garantizar la confiabilidad del instrumento, frecuentemente dejando a la validez como atributo subordinado y, la mayoría de las veces, sin citar a la objetividad como atributo indispensable. Para ilustrar esta afirmación, pero sin el propósito de realizar un meta análisis, la tabla 1 muestra la cantidad de entradas que se obtuvieron en un popular buscador de Internet para validez, confiabilidad, objetividad (en inglés y en castellano) y sus combinaciones. Los datos no representan preferencias definitivas de investigadores y evaluadores, pero ilustran la incidencia de los tres atributos en la Web. La objetividad es el atributo con menos referencias o entradas y, notablemente, su incidencia es menor en combinación con los otros atributos.

Tabla 1. Frecuencia de entradas dentro del buscador Google para "validity, reliability, objectivity"

	Frecuencia (en miles)	
Atributo	**Inglés**	**Castellano**
Validez	89,000	19,400
Confiabilidad	169,000	8,860
Objetividad	11,000	4,920
Validez y confiabilidad	4,187	243
Validez y objetividad	107.8	60.5
Confiabilidad y objetividad	183.7	32.2
Los tres atributos	26.1	2.15

Fuente: Elaboración propia a partir del buscador Google, octubre de 2016.

¿Cómo explicar que de 169 millones de entradas para confiabilidad (las cantidades son notablemente inferiores en castellano), se tenga una reducción a casi 4.2 millones al combinarse con validez y se tengan solo 26 mil entradas combinadamente? No parece explicable que un atributo tan importante se haya escapado a los especialistas en evaluación o a los investigadores de la psicometría.

Una explicación es que, de los tres atributos fundamentales de la evaluación y de la medición, la objetividad es el más complejo de definir y de aprehender. De hecho, varios de los principales factores de la objetividad se transfieren, erróneamente, a la validez, reduciendo a la objetividad a pocos aspectos, importantes pero insuficientes, castigando a la validez al contener factores que le son ajenos y que la convierten en un atributo altamente complejo; que además de referirse al concepto primigenio de que "el instrumento de medida sirva para el propósito previsto", también se asocia con la interpretación de los resultados. Así se han acuñado nuevos términos como "validez de uso", "validez consecuencial" y "validez cultural", entre otros, olvidando que están asociados con la objetividad. También la confiabilidad ha tenido que absorber aspectos que atañen a la objetividad, como es el sesgo de diseño.

Los propósitos de este trabajo son varios: aclarar el concepto de objetividad, explicar su importancia como uno de los tres atributos fundamentales de la evaluación y, sobre todo, establecer su papel dentro de las pruebas estandarizadas. Para cumplir con estos propósitos debe recurrirse a diversas áreas del conocimiento, pero se ha optado por abordar tres facetas teóricas para identificar las cualidades ontológicas, epistemológicas y éticas de la objetividad, incluyendo ejemplos en el terreno de la evaluación estandarizada.

Reflexionar sobre la objetividad en las pruebas estandarizadas no es trivial ni ocioso. Una vez comprendida su importancia, permite liberar de complicaciones a la validez y a la confiabilidad de las pruebas en general y de las estandarizadas en particular. Este trabajo no pretende hacer una reseña histórica de la objetividad a través de la filosofía y otras áreas del conocimiento; el lector interesado puede referirse al trabajo de Gaukroger (2012).

2. La objetividad como desiderátum

La objetividad es la cualidad inherente de un objeto en sí mismo, ajeno a cualquier enfoque especulativo (Real Academia Española, 2016; Zamora, 2007). Se ha utilizado en el tratamiento metódico y controlado para definir y estudiar "entes" y como base de discusión científica y filosófica al cuestionar la existencia "real" de las cosas (García, 1955) para determinar la posibilidad de alcanzar un conocimiento "real" del mundo fuera de otras aproximaciones. Las discusiones en torno a la objetividad han tratado de esclarecer su relación con la verdad, la realidad, la existencia y el ser como tal de los objetos, orientando el trabajo filosófico y científico desde el positivismo del siglo XIX hasta el relativismo contemporáneo, provocando diversas aproximaciones, tendencias y conflictos ontológicos.

La tensión entre objetivismo y subjetivismo ha permitido encontrar elementos útiles dentro de ambos, contribuyendo al desarrollo del conocimiento científico, particularmente en áreas predominantemente especulativas en el tratamiento del objeto de estudio, como en ciencias sociales y de la salud. Una postura más moderada reconoce las limitaciones de la visión objetiva mecánica, que pretende despersonalizar al observador para evitar que sus juicios afecten las descripciones que hace del objeto que analiza y, en cambio, reconocer que el observador no puede ser ajeno del todo a lo que observa, pero consciente de esta implicación y conociendo sus prejuicios, debe poder apartarlos del objeto en estudio (Cupani, 2011; Morales de Barbenza, 2001).

La objetividad es un desiderátum, es decir, inalcanzable plenamente por varias razones. Por una parte la ciencia, sus productos y motivaciones, son resultado de la actividad cognitiva que hace cada individuo sobre un objeto en particular; por otra parte, las representaciones o definiciones de un objeto están sujetas a la aprehensión consciente del investigador, lo cual queda obligatoriamente vinculado a su subjetividad al observar, medir, valorar, controlar o asignarle categorías lógicas dentro de un sistema teórico (Cupani, 2011). El estar consciente de esta limitante brinda la oportunidad de plantear aproximaciones al objeto por conocer, merced a un alto grado de "indiferencia en el juicio, que puede estar en conflicto con nuestras necesidades y deseos" (Gaukroger, 2012). En consecuencia el juicio se despersonaliza como si fuera hecho desde el exterior del propio sujeto.

Si se parte del argumento filosófico de que ningún objeto es aprehensible directamente, porque su existencia implica una declaración metafísica alrededor de cualidades inherentes u ontológicas, entonces, se acepta que todo objeto por conocer (real o ideal) posee un ser inteligible e identificable como correlato del objeto respecto de un conjunto de características específicas (García, 1955) que, al encontrar una expresión material (concreta o abstracta) adquieren realidad objetiva. El citado correlato puede ser resultado de una medición sobre el objeto, producto de un razonamiento deductivo formal o designarse por consenso de una comunidad científica o profesional, lo cual se denomina realidad subjetiva.

La objetividad es el resultado de un proceso dual que se basa, por una parte, en la contrastación de conocimientos e ideas en un mundo empírico y, por otra, en la intersubjetividad donde un grupo acepta la construcción de esa idea como válida por un acuerdo convencional sobre un mismo objeto partiendo de apreciaciones subjetivas. El proceso de objetivación reconoce las manifestaciones materiales de los objetos, independientes del observador, aceptando que, a pesar de que cada observador es único o singular, está en posibilidad de establecer criterios sobre sus afirmaciones, de modo que el acuerdo convencional elimina toda discrepancia entre observadores.

Los investigadores pueden definir formalmente, de manera conceptual o empírica, el objeto en estudio, en función de un contenido, una representación y una estructura con fundamento en cada área del conocimiento. El acuerdo convencional se fundamenta en definiciones dinámicas, aprovechando que la ciencia es autocrítica y se auto corrige, con estructuras que se construyen y reconstruyen al incorporar nuevo conocimiento proveniente de evidencia empírica o de postular nuevas categorías formales, en un ejercicio de honestidad intelectual (Gaukroger, 2012).

Una implicación evidente de no buscar sistemáticamente la objetividad como criterio científico en el sentido que le da Popper (citado por Larroyo, 1968), es que el conocimiento derivado de su estudio, puede no corresponder a las características o atributos del objeto o, peor aún, que los elementos estudiados sean plenamente dependientes del observador, haciendo que los atributos adjudicados al objeto estén más bien vinculados a otros procesos que no definen ni explican en absoluto lo que pasa con dicho objeto. En consecuencia, los atributos descritos por un observador serán discrepantes de lo que puede establecer otro observador, induciendo a que el tratamiento del objeto no sea sistemático y dependa de la interpretación de cada persona, de lo que se desprende la necesidad de la intersubjetividad citada previamente. Queda claro que la objetividad es un criterio fundamental en el desarrollo de la investigación científica, porque permite generar conocimientos válidos sobre los objetos investigados.

La objetividad depende de dos aspectos fundamentales: la especificidad y la interpretación.

1) La especificidad es la representación de la realidad, contenida en una definición completa, pertinente, precisa del objeto y que lo distingue de otros. Para esta definición se justifica el uso de un arquetipo como referencia para los juicios de valor que se pueden hacer de los objetos de su mismo género o especie. En consecuencia de su definición, la objetividad no es un constructo universal que todas las personas perciben de la misma forma, sino que requiere de la aceptación convencional a partir de las cualidades intrínsecas incluidas en la definición. La especificidad implica que la definición del objeto debe distinguir claramente entre

cualidades inherentes y otros elementos que pueden catalogarse como requisitos, criterios de inclusión o de exclusión, condiciones reglamentarias o administrativas para tener derecho a participar en un proceso de evaluación. Por ejemplo, el que la prueba PISA se administre a jóvenes de 15 años es un criterio de inclusión para el proyecto que restringe a otras personas a ser parte de la población focal, pero esta condición no se considera como un sesgo o una valoración subjetiva respecto de dicha población. Los requisitos no producen medidas respecto del objeto, por lo tanto no deben aportar calificaciones o puntajes a las personas de la población focal, ni tampoco generalizaciones que comprometan el uso ético de la información

2) La interpretación se asocia con las justificaciones de los usos y juicios de valor que pueden postularse a nivel contextual, cultural, grupal, o de otra índole, a partir de datos obtenidos de la realidad. Las interpretaciones y justificaciones responden a la necesidad de identificar, prevenir, medir y, de preferencia, eliminar o reducir al mínimo la presencia de sesgos en las apreciaciones de las personas que van a emitir juicios de valor sobre los objetos en estudio. Este tópico es complejo porque hay fuentes de sesgo imputables al proyecto, al evaluador y a la población, lo cual incide en problemas de diseño, al construir los ítems, al administrar la prueba, al emitir juicios de valor, de tal modo que las interpretaciones se ven afectadas por todas estas condiciones.

Al aceptar que la objetivación es factible de ser alcanzada y definida, se avanza contra el escepticismo que niega dicha factibilidad, en particular en las ciencias sociales, psicología, educación y áreas de la salud; aunque, en el extremo, la objetividad matemática tampoco sería alcanzada por tratar con entes abstractos cuya manifestación real es siempre imperfecta. Por ejemplo, el concepto de triángulo como figura plana cerrada de tres lados, cuyos ángulos internos suman 180º es geométricamente perfecto, pero solo puede dibujarse en un papel de manera aproximada por un dibujante experto. Sin embargo, el arquetipo del triángulo es una formalización que perfecciona la percepción que se hace de un objeto real que solo puede existir de manera imperfecta; este perfeccionamiento lo hace objetivo y permite que se interpreten sus propiedades de la misma manera por el dibujante, el ingeniero que dirige la edificación, el albañil que ejecuta la construcción o la persona que va a contemplar la obra terminada. De igual modo, en el campo de la evaluación, se tiene que mirar la objetividad de los modelos estadísticos como paradigmas esperados del comportamiento de un ítem o de un test estandarizado, lo cual invalida las objeciones que rechazan la construcción de modelos teóricos al centrarse en lo que denominan "evidencia empírica", dejando abierta la relatividad de la recolección de los datos, el juicio del observador y cualquier otra fuente de subjetividad en la definición del constructo, el diseño del instrumento y la interpretación de los resultados. Como apuntan Myers y Hansen (2002), puede afirmarse que la objetividad no niega la utilidad de recabar datos de la realidad, pero advierte que no son suficientes para garantizar que se obtienen conclusiones correctas.

Algunas cualidades que posee la objetividad se pueden asimilar a propiedades asociadas con ella (Gaukroger, 2012), en particular (1) la ausencia de sesgo en la interpretación y la toma de decisiones, (2) la eliminación de prejuicios personales, por lo tanto, libres de supuestos y valores individuales, (3) la facultad de distinguir entre dos ideas o teorías contrastantes o hasta en conflicto respecto del objeto, sustentándose en (4) la definición exacta del objeto. Dicho autor advierte en no asociar la objetividad exclusivamente con la cuantificación y la medida o con la acumulación e interpretación de datos. La medida

es parte importante de la objetivación, pero no es el único elemento que la constituye, porque reduciría el concepto de estandarización a la obtención de resultados que pueden medirse, analizarse en forma matemática o estadística para su comparación.

Entrando en el terreno de la evaluación y como consecuencia de lo expresado, la definición del objeto es el punto de partida para el proceso de evaluación porque es el que permite hacer apreciaciones cualitativas o cuantitativas respecto de los atributos inherentes del objeto, propiciando profundizar en sus características y funciones. La definición del objeto a medir debería ser una de las responsabilidades y preocupaciones de los diseñadores de pruebas y cuestionarios, porque sin una definición objetiva, es altamente probable que se obtengan resultados y conclusiones poco fieles del objeto que, en este caso, se trata de cualidades de las personas evaluadas; donde el resultado de esa medición tendrá un impacto muy importante en su vida; como en el caso de ser admitido o rechazado en la universidad; del diseño de políticas y programas gubernamentales; del establecimiento de una campaña de salud pública. De no hacer una adecuada definición puede verse el enorme riesgo de generar un instrumento que proporcione información deficiente en términos de objetividad.

La evaluación depende de la objetivación para definir el objeto de medida. Por ejemplo, puede tenerse interés de disponer de un instrumento para medir la temperatura corporal, lo cual se resuelve fácilmente con adquirir un termómetro en la farmacia más cercana. Este aparato debe ser válido y confiable para obtener medidas certeras de temperatura al utilizarlo con cualquier persona. Todo parece simple, siempre y cuando el concepto de temperatura tenga una definición objetiva, sin confundir "temperatura" con "calor", "fiebre", "dilatación" u otro concepto con el cual esté posiblemente asociada la temperatura, pero que se manifiesta, mide e interpreta de otra forma. El funcionamiento del termómetro será válido para el propósito de medir el objeto deseado (temperatura), sin cuya definición sería inapropiado el uso del instrumento e inútiles las medidas resultantes. La objetividad, como desiderátum precede, por lo tanto, a la validez y a la confiabilidad.

3. La objetividad como herramienta epistemológica

Una forma de aproximarse a un objeto concreto es identificar algunas propiedades físicas observables: dimensiones geométricas, características de materiales y forma, cantidad de un atributo como peso o temperatura (Nunnally y Bernstein, 1995). Los investigadores al medir buscan "asignar símbolos a objetos de manera que (1) representen cantidades o atributos de forma numérica (escala de medición) o (2) definan si los objetos caen en las mismas categorías o en otras diferentes con respecto a un atributo determinado (clasificación)" (Nunnally y Bernstein, 1995). La forma de identificar las propiedades puede ser tanto teórica como experimental, por lo que se han ideado instrumentos y estrategias para medir uno o varios atributos del objeto en escalas apropiadas.

La medición de un objeto abstracto (como la inteligencia, la percepción de un síntoma hepático, la depresión, el aprendizaje, entre otros), presenta importantes limitaciones prácticas, ya que la relación existente entre el objeto y su realidad objetiva no es directa, como en los objetos concretos (Kerlinger y Howard, 2008b). Para el análisis y la medición los investigadores precisan definir un conjunto de características del objeto, concentrándose en una o en algunas manifestaciones de ellas, denominadas "rasgo" (Nunnally y Bernstein, 1995). En el caso particular de los rasgos observables de forma

indirecta a través de comportamientos y expresiones diversas que realice una persona, se habla de rasgos latentes y se presume que de forma indirecta pueden ser medidos al observar tales manifestaciones que contienen la característica prevista del objeto. La forma de vincular la realidad objetiva y el rasgo latente es altamente compleja, con gran probabilidad de confusión e imprecisión, máxime que distintos enfoques científicos, áreas del conocimiento o sistemas teóricos pueden estudiar simultáneamente el mismo rasgo latente atribuyéndole propiedades, funciones y manifestaciones distintas.

Al definir el objeto de medida en un proyecto de evaluación se pueden identificar los límites, alcances, interpretación y uso de los resultados (Jornet y Suárez, 1996). De nuevo, la objetividad precede a los atributos de validez y confiabilidad, dirigiéndolos en el rumbo previsto por el proyecto evaluativo, de tal modo que las consecuencias de una evaluación no son inherentes a estos dos atributos, sino a la objetividad que las antecede. De este modo la objetividad no se limita a la definición del objeto de medida y a su interpretación, sino también contiene un conjunto mucho más amplio de propósitos, entre ellos:

a) Aprehender las cualidades inherentes del objeto, al definirlo, caracterizarlo, categorizarlo, compararlo, ponderarlo, valorarlo o medirlo, entre otras formas.

b) Emitir juicios de valor sobre uno o varios rasgos o características inherentes del objeto, en función del objeto mismo, de una población dada, o respecto de criterios externos de referencia o de comparación.

c) Plasmar (en forma conceptual, simbólica, matemática o de otra índole) las cualidades, características o rasgos de un objeto para su análisis y aprehensión por diversas personas, incluyendo el evaluador y el evaluado, o un público independiente.

d) Reducir el sesgo de diseño del instrumento para propiciar la apreciación formal del objeto con especificaciones definidas en forma concreta u operacional, en forma independiente de las poblaciones en las que se utilice.

e) Acotar la interpretación subjetiva del evaluador respecto del rasgo evaluado, en un momento dado, o a lo largo del tiempo por cambios de criterios que experimenta el evaluador.

f) Reducir la diferencia de apreciación de diversos evaluadores, en función de criterios, consideraciones o prejuicios personales.

g) Evitar la diferencia de apreciación entre el evaluador y el evaluado, haciendo que este último perciba su dictamen como aceptable.

h) Eliminar el efecto de fatiga o influencia cualitativa por el número de juicios emitidos en un tiempo dado ante una población numerosa.

i) Anular el efecto de halo, de prejuicios discriminatorios o por influencia de estereotipos en el evaluador.

j) Eliminar la discrepancia de opinión respecto de la respuesta correcta o más aceptable, facilitando la calificación por personal no experto e, inclusive, por medio de un programa informático con base en una clave de respuestas.

k) Comparar las cualidades métricas de varios instrumentos, incluyendo el error de medida y la consistencia de resultados que se obtienen con una población focal dada.

l) Obtener medidas de los ítems independientemente de la población o personas particulares que intervienen en la aplicación del instrumento y, en contraparte, obtener medidas de las personas de la población focal independientes del conjunto de ítems utilizados en el instrumento.

Las pruebas estandarizadas son los instrumentos de medición más utilizados en psicología, educación, ciencias de la salud y ciencias sociales, que cuentan con un amplio desarrollo técnico y metodológico con formas perfeccionadas para medir los rasgos observables o latentes, en la población focal específica y con un grado de precisión previamente establecido y controlado por procedimientos logísticos y administrativos igualmente objetivos. Los atributos de validez y confiabilidad de las pruebas estandarizadas han sido objeto de muchos debates y de críticas que no se repetirán en este trabajo, sin embargo, parte de ellas se deben al limitado, por no decir nulo papel que se le ha dado a la objetividad como atributo de las pruebas estandarizadas (Borsboom, Mellenbergh y Heerden, 2004; Embretson, 2007; Kane, 2008; Kerlinger y Howard, 2008a-b; Mislevy, 2007; Newton y Baird, 2016; Padilla, Gómez, Hidalgo y Muñiz, 2006; Sijtsma, 2009).

La idea de base de las pruebas estandarizadas como instrumentos de medidas de objetos abstractos o rasgos latentes cuenta con una profunda influencia del positivismo del siglo XIX, que buscaba establecer con el mayor rigor metodológico posible una definición del objeto de estudio, por ejemplo, la inteligencia o el rendimiento escolar (Binet, 1910), asumiendo que las manifestaciones observables de los rasgos latentes son objetivaciones que se quieren medir en el objeto y cuyos resultados se reportan en una escala es un eje cartesiano igualmente objetivo, que corre de menos a más respecto del atributo. Para garantizar la precisión de los resultados se cuenta con medidas objetivas del error que requieren de un control también objetivo de las situaciones en las que realiza la medición, todo lo cual anula o, por lo menos, reduce la influencia de variables que afecten la medición, por ser dependientes de varios agentes: el evaluador en el diseño del instrumento objetivo, el aplicador al administrar la prueba de forma objetiva y de la persona a evaluar al responder ítems objetivos por medio de un desempeño objetivo (Binet, 1910). Todos los elementos fueron adjetivados con la palabra "objetividad", recordando la necesidad de que la prueba sea objetiva, pero no debe pensarse que todas las pruebas objetivas están estandarizadas y, desde luego, no puede garantizarse que todas las pruebas estandarizadas disponibles en el mercado sean objetivas.

La objetividad sirve al propósito de la medición, ayuda a definir y delimitar el objeto a evaluar, así como a proporcionar elementos de control de dicha medición, para limitar que variables externas afecten el resultado, y que el medio ambiente, incluyendo al administrador de la prueba, no interfiera con el resultado.

Los valores y los propósitos de la objetividad contribuyen a reducir o constreñir la intervención subjetiva de quien administra o evalúa una medición (Cupani, 2011), reduce la influencia del evaluador cuando corrige una prueba, al valorar el resultado final, el nivel de desempeño de un estudiante en una asignatura escolar o el grado de enfermedad de un paciente ante un síntoma (Céspedes, 2009). De nada sirve contar con una prueba estandarizada de buena calidad, si la persona que va a aplicarla e interpretar los

resultados no está capacitada o si el dictamen final depende de criterios no ligados al objeto. Por lo tanto, la objetividad incide en varias fases del proyecto de evaluación: la planeación y diseño del instrumento, su administración, el control del proceso y la logística, la calificación y la interpretación, por ello no solo precede a los atributos de validez y confiabilidad como se indicó en la sección anterior, sino que funciona como un control de calidad de cada etapa de desarrollo de ambos atributos.

Puede verse, por lo tanto, que la definición primigenia de validez como el grado en que una prueba mide el propósito que se pretende medir es muy apropiada, porque asume que el objeto de medida fue definido claramente (Kelley, 1927). Además, cada etapa que permite obtener evidencias de validez se concentra en la sensibilidad del instrumento para captar el objeto y los atributos definidos en su objetivación. Este reconocimiento hace evidente que toda medida es imperfecta y como tal, tiene un margen de error que se vincula y calcula a través de procedimientos estadísticos, que objetivan a la confiabilidad. La definición de estos atributos, entre muchos otros, ha sido emprendida por diversas agencias o instituciones (APA, 1954-2010; AERA, APA, NCME, 2014), las cuales han sido sometidas a análisis, críticas y escrutinios dentro de la comunidad académica (Campbell, 1960; Chan, 2014; Guilford, 1987; Jeffrey, 2003; Kimberlin y Winterstein, 2008; Lane, 1999; Moss, 2007; Newton y Baird, 2016; Sireci, 2007; Sireci y Padilla, 2014).

La definición primigenia de validez es objetiva en los conceptos de validez de contenido, de constructo, de criterio (predictiva, concurrente, discriminante…) y de escala, pero se modificó el modelo al plantearse que la validez no es un atributo inherente del instrumento sino que depende del uso e interpretación que se haga de los resultados, lo cual involucra implicaciones éticas (Borsboom, Mellenbergh y Heerden, 2004; Chan, 2014; Jeffrey, 2003; Lissittz y Samuelsen, 2007; Messick, 1995; Zumbo, 2009). De esta forma, el uso y la interpretación caen en el terreno de la objetividad, no siendo pertinente adjudicarlos a la validez, porque esto complica y enturbia su significado dentro de la evaluación y despoja a la objetividad de algunos de sus propósitos.

Respecto de la posible confusión entre objetividad y validez, es importante citar que, de acuerdo con Borsboom et al. (2004), una prueba es válida cuando el atributo existe y sus variaciones producen causalmente variaciones en la medición. Esta definición de validez, parece un sano retorno al concepto inicial pero con base en un sustrato distinto, al surgir de una reflexión ontológica (André y Loye, 2015; Jeffrey, 2003) sobre la objetivación de "aquello" que se quiere medir, distinguiendo los rasgos inherentes al objeto de los que no lo son. Si un objeto cambia, entonces se debe reflejar un cambio en su medida, lo que requiere de un proceso constante de objetivación y mantener esa vigilancia durante el proceso de medición. En caso contrario, es indispensable objetivar nuevamente el objeto y su medida, lo cual puede repetirse las veces que sean necesarias para garantizar que las medidas y las unidades que se utilizan miden lo que deben medir. Aceptando que la objetividad es el sustrato de la validez, en ausencia de ella, la validez queda seriamente comprometida.

Una prueba estandarizada debe tener claramente objetivado el rasgo con elementos de la realidad objetiva y de la realidad subjetiva. Para operacionalizarlo es posible utilizar enunciados, categorías y variables susceptibles de ser exploradas de forma cualitativa o cuantitativa. Todas las pruebas, en particular las estandarizadas, deberían usar diversas técnicas para comprobar que la operacionalización corresponde a los rasgos que se pretende medir. Esta comprobación puede hacerse a través del consenso del juicio de

expertos (evaluación de realidad subjetiva por terceros), con pruebas de correlación entre ítems, ítem contra prueba, entre pruebas distintas, con la misma prueba a lo largo del tiempo o con poblaciones de contraste, entre muchas otras formas.

En los propósitos de la evaluación objetiva se asocia la operacionalización con la independencia entre el evaluador y el evaluado, entre la medida del ítem y la del sujeto. La independencia es una cualidad de la objetividad que sistematizó Rasch (1980) con el concepto de independencia local y que garantiza que la probabilidad de respuesta de un sujeto ante un estímulo dado es una función que depende de la medida del sujeto y de la dificultad del ítem, independientes entre sí. Este modelo se ha extendido al análisis de facetas múltiples que permite incluir la opinión de los evaluadores y de variables de contexto (Linacre, 1994).

En general la confiabilidad ha tenido menos conflictos de interpretación que la validez, especialmente si se toma en el sentido de expresar valores relacionados con el grado de precisión de las medidas (Nunnally y Bernstein, 1995), pudiendo provenir de modelos que estiman la consistencia de los datos, la homogeneidad de los ítems y de la población, o la repetitividad de los resultados cuando la prueba es administrada a lo sujetos en condiciones controladas (Argibay, 2006; Carvajal-Carrascal, 2012; Kerlinger y Howard, 2008a; Sánchez-Meca, López-Pina y López, 2009; Zúñiga y Montero, 2007), siendo el Alfa de Cronbach, la teoría G y la separación logística, los modelos más utilizados en la práctica, dentro de un abanico enorme de modelos que persiguen calcular el error de medida de cada ítem, de la prueba en su conjunto, de los puntos de corte, entre otros elementos que tratan de brindar medidas objetivas de la precisión de la medida, aunque no de la calidad del instrumento. Tradicionalmente, los valores aceptables del Alfa de Cronbach se dejan a juicio del evaluador, es decir, quedan supeditados a criterios subjetivos (Blanco-Villaseñor, 1991; Nunnally y Bernstein, 1995) por lo que no se ve problema en aceptar un valor de Alfa de 0.8 en una prueba estandarizada y se rechaza que una de las partes de la prueba tenga valores tan bajos como 0.4 (Tristán, 1996-2010). Es posible establecer criterios objetivos para demostrar la pertinencia de ambos valores sin apelar a artificios en el diseño (incrementar el número de ítems o restringir la dificultad de los ítems alrededor del punto de corte) conduciendo a un instrumento con una alta confiabilidad a expensas de una pobre validez.

Modelos matemáticos y estadísticos más sofisticados favorecen la creación de herramientas que incorporan distintos supuestos sobre las variaciones en las puntuaciones (Shavelson y Webb, 2005; Ritter, 2010) en particular a través de modelos logísticos o multivariados para analizar el funcionamiento diferencial de cada ítem o de la prueba en su conjunto, con énfasis en reducir o corregir el sesgo inherente al diseño o relativo a la población evaluada (Bond y Fox, 2015; Fox y Glas, 2001,2003; Gómez y Hidalgo, 2003; Jiménez y Montero, 2013; Linacre y Wright, 1995; Prieto y Delgado, 2003; Wright y Stone, 1999; Wright y Mok, 2000). Tomar en cuenta el funcionamiento diferencial o la presencia de algún sesgo es fundamental al emitir juicios de valor sobre personas en forma individual o grupal, lo cual va más allá del interés estadístico por sus consecuencias éticas.

4. Objetividad y consideraciones éticas en las pruebas estandarizadas

El método científico tiene como característica inmanente (explícita o no) a la objetividad (Muñiz, 2010), porque se espera que las preferencias, actitudes, valores y prejuicios del investigador no afecten su trabajo. Se extrapola esta idea a las pruebas estandarizadas, al desarrollar instrumentos de medición en las ciencias sociales y de la salud perfeccionados con técnicas psicométricas y predictivas con rigor científico. Este desarrollo diluyó aparentemente la discusión sobre la relevancia, la utilidad y las implicaciones del uso ético de las pruebas (André y Loye, 2015), en parte por el tiempo que ha implicado desarrollar técnicas y software de análisis estadístico, así como enfrentar cierto rechazo a las pruebas estandarizadas, a la pertinencia de su uso y puesta a disposición de profesionales certificados para su administrarlas, interpretar los resultados y tomar decisiones prácticas dentro de un marco ético o de justicia para las personas evaluadas.

Los artículos de difusión de resultados, especialmente los de la segunda mitad del siglo XX en los Estados Unidos de América, trataban de convencer al lector de los beneficios de la estandarización desde el punto de vista positivista, vinculando el desempeño (intelectual, académico y laboral) con grupos de personas, mostrando diferencias entre géneros, etnias, culturas y niveles socioeconómicos, reforzando estereotipos y clasificaciones discriminatorias (Herrenstein y Murray, 1994; Bowen y Bok, 1998), provocando un impacto político y social resultante de algunas debilidades de estas pruebas. Las soluciones se concretaron de varias maneras: La primera fue criticando los defectos de las pruebas, promoviendo su erradicación en el ámbito de la educación y sugiriendo modelos de evaluación "auténtica" (Froese-Germain, 1999). Una segunda línea fue de tipo legal bajo sentencias judiciales y enmiendas del Congreso de los Estados Unidos (Enmienda Buckley de 1976 o FERPA) para supeditar el papel de las pruebas estandarizadas a los derechos civiles, durante la aplicación, la calificación y la utilización de los tests (Gómez, Hidalgo y Guilera, 2010; Nunnally y Bernstein, 1995). La tercera línea técnica construyó estándares para el diseño de pruebas por el Joint Commitee (AERA-APA-NCME, 2014), o estándares de buenas prácticas y equidad en las pruebas (Educational Testing Service, 1987; International Test Commission, 2014-2016). Una cuarta línea defendió las pruebas estandarizadas con base en argumentos objetivos, (curiosamente sin invocar a la objetividad) contrastando sus ventajas contra otras formas de evaluación (Phelps, 2005).

La defensa de las pruebas estandarizadas ha implicado aportar elementos para corregir deficiencias reveladas por las críticas de sus detractores con un impacto ético. Estos elementos agregados sobre todo a la validez y a la confiabilidad las convierten en atributos "ómnibus" que absorben todo lo que permita reforzar a las pruebas, pensando que enderezan el camino de las pruebas estandarizadas pero que enturbian su existencia, complicando su vulnerabilidad en el campo ético frente a una mirada inquisitiva y crítica. Toda proporción guardada, son empeños similares a los que defendían el modelo geocéntrico de Tolomeo, agregando elementos complicados y tortuosos para explicar la cinemática de los cuerpos celestes, frente al modelo heliocéntrico de Copérnico, simple, claro y preciso. Las implicaciones éticas de la objetividad se relacionan con las propiedades de neutralidad, imparcialidad e impersonalidad del observador-evaluador.

La impersonalidad hace explícitas y conscientes las representaciones culturales y sociales implicadas en una prueba estandarizada y, por lo tanto, bajo la responsabilidad

de las personas que la desarrollan, desde los consejeros que determinan el objeto de medida, hasta los responsables de su utilización e interpretación, pasando por los diseñadores de ítems y los encargados del procesamiento estadístico. Es fundamental definir claramente el objeto de medida, sus interacciones con factores psicológicos, biológicos, ambientales y de experiencias previas que puedan afectar o condicionar la obtención de evidencias sobre el objeto, especialmente cuando es un rasgo latente. La representación debe explicitar cómo el objeto es compartido en el grupo social, cultural, étnico, en un momento dado o en su devenir temporal y contextual (etario, regional, socioeconómico). La impersonalidad obliga a adaptar una prueba creada en un idioma o país para aplicarse en otro, no solamente como traducción sino como concepción del objeto, definiendo las situaciones o casos que describen y aclarándolas para cada contexto. Esto requiere de un arduo trabajo de interpretación de la prueba, de validación para cada población y el establecimiento de criterios de corte y baremos para los diversos grupos poblacionales (Muñiz, Elosua y Hambleton, 2013; Sattler, 2010).

La neutralidad requiere que no haya injerencia externa en los juicios de valor que emite un evaluador con los resultados de una prueba estandarizada, haciéndola aplicable a todas las personas, en todos los ambientes y condiciones, obteniendo medidas libres de otras características ajenas al objeto. Por ejemplo, se tiene un problema de neutralidad en una prueba aplicada por un sindicato para clasificar personal en un puesto de trabajo, si el resultado que se emite es distinto cuando las personas están sindicalizadas o no. En el caso de la prueba PISA se tiene un problema de falta de neutralidad, si los textos utilizados como situación para derivar los ítems hacen referencia a objetos comunes en un país y que no son comprensibles para los estudiantes de otro.

Una prueba de comprensión lectora sobre el tópico central de un texto y diversos aspectos gramaticales concibe que ambos son constructos neutrales y no personalizados. De hecho, se puede plantear sobre un texto que describa la belleza del campo (neutral y no personalizada), o sobre un texto que detalle una situación de violencia social (personaliza aunque puede ser neutral si no toma una postura) o un relato que ridiculice a los seguidores de una religión (personaliza y no es neutral por demeritar al grupo en cuestión). La respuesta ante esos estímulos será diferente porque movilizará en cada persona sentimientos y reacciones ajenas al propósito de medida.

La imparcialidad pretende garantizar que la prueba estandarizada sea justa, sin prejuicios ni sesgos (Gómez, Hidalgo y Guilera, 2010), de tal modo que las medidas que se obtienen de ella sean resultado de la comparación de un rasgo en condiciones de equidad contextual (Nunnally y Bernstein, 1995). El análisis de imparcialidad o carencia de sesgo, hace indispensable el reconocimiento escrupuloso de todas las variables que pueden inducir a respuestas no objetivas, con las que se producen medidas erróneas y apreciaciones injustas a personas de un grupo específico, en función de género, grupo etario, nivel socioeconómico, antecedentes culturales, pertenencia religiosa o étnica, entre otras. En ese sentido, los investigadores deben cuidar que el lenguaje, las situaciones y el contexto de los ítems no vulneren la dignidad de las personas, que no induzcan la movilización de rasgos latentes no previstos que pudieran favorecer que se movilicen actitudes positivas o negativas en ciertos grupos o individuos.

El análisis de sesgo debe hacerse a priori, al definir el objeto y las especificaciones de diseño de la prueba y a posteriori con técnicas estadísticas avanzadas para detectarlo, medirlo y realizar ajustes matemáticos de cambio de escala e igualación de los resultados obtenidos por los grupos potencialmente afectados por dicho sesgo. Es muy

acostumbrado entrar en un proceso tautológico utilizando un discurso subjetivo para explicar la falta de imparcialidad con base en valores de comparación o puntos de corte sin justificación objetiva, haciendo que las conclusiones estén igualmente sesgadas y, por lo tanto, carezcan también de imparcialidad.

Al ignorar que la objetividad requiere satisfacer estas propiedades se transfiere el problema a decidir si es válido utilizar un instrumento para fines distintos a los que motivan su diseño, si los resultados son válidos para determinado grupo, o si es válido hacer dictaminar a un individuo con los resultados de una prueba independientemente de sus consecuencias. Obsérvese que se acostumbra usar coloquialmente la palabra "válido" pero no en el sentido estricto de "validez", con lo que se confunden los propósitos y conceptos, haciendo que la validez -y no la objetividades- se asocie con el contexto cultural, con los usos y las consecuencias de la interpretación de los resultados (Messick, 1993-1995; Prieto y Delgado, 2003). Es de esperarse que la triada objetividad-validez-confiabilidad oriente el interés de los evaluadores hacia las implicaciones éticas, de equidad y de justicia. Como apuntan Kovač-Šebart y Krek (2009): "objetividad, validez y confiabilidad son categorías interconectadas e interdependientes, y todas ellas están incluidas en la percepción de la justicia".

5. Conclusiones

La objetividad incide, como se ha visto, en todos los factores y las etapas de la evaluación en general y en el desarrollo de una prueba estandarizada en particular. Puede decirse que, junto con la validez y la confiabilidad, forma una cadena interactiva, donde intervienen simultáneamente. Sin embargo debido a la necesidad de definir objetivamente el objeto de medida como primer elemento en el proceso de evaluación y como auxiliar en el desarrollo de la prueba, la objetividad es el primero de los atributos, solo a partir de ella es posible cuestionar si el instrumento es válido y confiable.

La objetividad debe verse como una brújula que orienta el desarrollo de un proyecto de evaluación, siendo al mismo tiempo la línea de horizonte hacia la cual debe caminarse de forma continua, debido a que es la única manera de garantizar que se cumple con los propósitos científicos de las pruebas estandarizadas. Negar la objetividad o relegarla a una posición diferente a ésta, genera confusión y ambigüedad en el desarrollo de una prueba, redundando en medidas con una validez potencialmente dudosa y una confiabilidad de interpretación poco clara, además de contribuir a configurar un contexto que puede incidir en uso inadecuado y poco ético de los resultados.

Las propiedades que resultan de los tres ejes teóricos utilizados en este trabajo permiten identificar los elementos indispensables de la objetividad, con ellos se puede llevar a cabo una vigilancia práctica en cada etapa del desarrollo de una prueba estandarizada. La tabla 2 incluye un ejemplo correspondiente a una prueba olímpica (patinaje artístico) que el lector podrá adaptar a otras aplicaciones.

Tabla 2. Propiedades de la objetividad en las pruebas estandarizadas (I)

PROPIEDAD	**1. ESPECIFICIDAD**
La prueba tiene este atributo si:	*Cuenta con una definición completa, pertinente, precisa del objeto, que lo distingue de otros*
Propósito en las pruebas estandarizadas	**Ejemplo**
1.1 Definir el objeto, modelo de medición, registro de los rasgos, análisis de datos y resultados del instrumento para que no se vean influidos por cualidades ajenas al objeto mismo. La aprehensión del objeto debe ser hecha con base en cualidades inherentes, en función de sus características, categorías, comparaciones, ponderaciones, valoraciones o medidas y arquetipos, entre otras formas.	Fuera de los aspectos reglamentarios y de la organización por categorías, la calificación debe hacerse con criterios asociados a la ejecución artística (belleza, gracia, estética de movimiento…) y los aspectos técnicos (cualidades de la carrera de frente, de espaldas, de los saltos…), pero no debe considerar nacionalidad, religión, grupo étnico o edad de los patinadores como criterio para ser asignada.
1.2 Distinguir claramente entre dos ideas contrastantes o hasta en conflicto respecto del objeto.	Dos jueces pueden explicar y justificar las calificaciones respecto de un patinador, reconociendo sus aciertos o errores.
1.3 Distinguir entre las características inherentes medibles del objeto y los requisitos no medibles construidos alrededor del mismo.	El reglamento establece claramente las categorías por género o por tipo de discapacidad para las competencias de patinaje.
1.4 Comparar las cualidades métricas de varios instrumentos, incluyendo el error de medida y la consistencia de resultados que se obtienen con una población focal dada.	Un modelo de facetas múltiples puede brindar medidas de habilidad de los patinadores en diversas ejecuciones de dificultad dada, de la severidad de los jueces y del error de medida de cada caso.

Fuente: Elaboración propia.

Tabla 3. Propiedades de la objetividad en las pruebas estandarizadas (II)

PROPIEDAD	**2. NEUTRALIDAD**
La prueba tiene este atributo si:	*No hay injerencia externa en los juicios de valor que hace un evaluador u otras personas con los resultados de una prueba estandarizada.*
Propósito en las pruebas estandarizadas	**Ejemplo**
2.1 Reducir o evitar la interpretación subjetiva del evaluador en un momento dado o a lo largo del tiempo, inducida por la fatiga o el número de juicios emitidos en una población numerosa).	El juez dispone de criterios para asignar calificaciones iguales al principio y al final de la competencia, comparables con calificaciones de otros patinadores en eventos previos.
2.2 Evitar o reducir la diferencia de apreciación entre dos evaluadores o entre el evaluador y el evaluado.	Las discrepancias entre jueces ante el desempeño de un patinador deben reducirse al mínimo. El patinador y su entrenador (u otra persona experta) deben percibir que la calificación emitida no difiere de lo que ellos mismos pueden juzgar.
2.3 Evitar que grupos específicos puedan verse favorecidas o perjudicadas por el diseño de la prueba o la apreciación del evaluador.	Un juez califica de forma más benévola a los patinadores de su mismo país para ayudarlos. Otro juez es más severo con los patinadores de su país para evitar que piensen que hace favoritismo.
2.4 Eliminar la discrepancia de opinión respecto de lo que se considera la respuesta correcta o la más aceptable, facilitando la calificación por personal no experto o por medio de un programa informático.	Las puntuaciones emitidas por los jueces deben ser verificables dentro de su orden de error. El público (persona no experta) puede reconocer que la calificación del patinador es aceptable siguiendo los mismos criterios y emitir calificaciones equiparables.

Fuente: Elaboración propia.

Tabla 4. Propiedades de la objetividad en las pruebas estandarizadas (III)

PROPIEDAD	**3. INDEPENDENCIA**
La prueba tiene este atributo si:	*Las medidas y juicios de valor no se ven influidas por otros rasgos, instrumentos o agentes, personales o contextuales.*
Propósito en las pruebas estandarizadas	**Ejemplo**
3.1 Permitir que la medida de cada persona no se vea influida por las medidas de las otras personas a las que se administra la prueba, ni tampoco por las características propias del instrumento utilizado.	Las calificaciones de los patinadores no deben darse en comparación con otro patinador sino respecto de los atributos de su desempeño.
3.2 Favorecer que la medida de cada ítem no se influya por las medidas de otros ítems incluidos en el instrumento, ni por las características de grupos específicos en los que se administra la prueba.	Las calificaciones de los desempeños artístico y técnico del patinador deben ser independientes entre sí.
3.3 Garantizar que el juicio que emite un evaluador no refleje la influencia u opinión de otro evaluador.	Cada juez emite la calificación del patinador sin ver las de los otros jueces.
3.4 Garantizar que el juicio que emite un evaluador no se vea influido por datos previos de cualidades del sujeto o del conjunto de personas a evaluar.	Cada patinador debe ser calificado sin tomar en cuenta su desempeño en un evento anterior.

Fuente: Elaboración propia.

Tabla 5. Propiedades de la objetividad en las pruebas estandarizadas (IV)

PROPIEDAD	**4. INDEPENDENCIA**
La prueba tiene este atributo si:	*Las medidas y juicios de valor no se ven influidas por otros rasgos, instrumentos o agentes, personales o contextuales.*
Propósito en las pruebas estandarizadas	**Ejemplo**
4.1 Emitir juicios de valor libres de sesgo sobre uno o varios rasgos o características inherentes del objeto mismo.	Los jueces emiten su calificación basados en el desempeño de los patinadores sin importar su género, país de procedencia, pertenencia étnica u otro aspecto ajeno al patinaje.
4.2 Eliminar en el evaluador el efecto de halo, de prejuicios o estereotipos.	El juez emite una calificación más favorable a los patinadores procedentes de países con mayor tradición en esta disciplina.
4.3 Otorgar a todas las personas evaluadas las mismas oportunidades para mostrar su desempeño ante un instrumento dado, previas adaptaciones por discapacidades u otra característica justificada.	Las reglas para calificar los elementos de una rutina de patinaje de pareja deben ser las mismas independientemente del género de los patinadores.

Fuente: Elaboración propia.

Tabla 6. Propiedades de la objetividad en las pruebas estandarizadas (V)

PROPIEDAD	**5. IMPERSONALIDAD**
La prueba tiene este atributo si:	*Explicita la forma en que el objeto es compartido en el grupo social, cultural, étnico u otro al que pertenece en un momento dado, considerando su evolución en el tiempo y en cada contexto.*
Propósito en las pruebas estandarizadas	**Ejemplo**
5.1 Evitar que personas específicas puedan verse favorecidas o perjudicadas en la prueba.	El juez no emite su calificación a partir de la trayectoria deportiva del patinador sino sobre el desempeño concreto observado.
5.2 Plasmar las características o rasgos de un objeto transparentando su análisis y aprehensión por diversas personas, incluyendo el evaluador y el evaluado, o un público independiente.	La apreciación del juez sobre las características técnicas de las piruetas está plenamente descrita en las reglas disponibles por el patinador, su entrenador y los diferentes jueces.
5.3 Validar los usos e interpretaciones a nivel contextual, cultural, grupal, o de otra índole, que se postulen a partir de datos obtenidos de la realidad.	La apreciación del juez sobre las características técnicas de una pirueta no debe verse modificada en función del origen étnico del patinador.

Fuente: Elaboración propia.

Incorporar la objetividad como atributo principal del proceso de evaluación es particularmente imprescindible en educación y ciencias sociales, no solamente para definir objeto a evaluar, sino por el uso de las pruebas estandarizadas de selección para ingreso a universidad o certificación profesional. Pocas veces se cita la objetividad junto con validez y confiabilidad en las pruebas estandarizadas, proliferando los detractores que objetan que sean "válidas" para evaluar a los estudiantes de ambiente rural, de etnias monolingües que no dominan el idioma nacional o que pertenecen a zonas deprimidas del país, sobre la base de que están en desventaja respecto de los estudiantes urbanos y de alto nivel socioeconómico, haciendo que la interpretación de sus resultados tenga implicaciones y consecuencias negativas para ellos. Debe quedar claro, por lo tanto, que no se trata de un asunto que pueda resolver la validez sino la objetividad, porque al usar una prueba en toda la población focal se tiene la ventaja establecer comparativos útiles para las políticas educativas y sociales del país, así como hacer interpretaciones diferenciadas entre grupos poblacionales.

La prueba PISA, promovida por la Organización para la Cooperación y el Desarrollo Económicos (OCDE), cumple con altos criterios de validez y de confiabilidad, pero su objetividad es cuestionable debido a que, fuera de que usa ítems objetivos, no hace explicita su relación con este atributo. Entre las versiones de 2003 a 2015 (OECD, 2005-2016), solo se menciona en dos reportes nacionales (Eslovaquia y República Checa) vinculándola con la neutralidad y la imparcialidad para garantizar medidas objetivas sobre el desempeño (Santiago, Halász, Levacíc y Shewbridge, 2016; Shewbridge, Herczyński, Radinger y Sonnemann, 2016).

Alcanzar la objetividad en el proceso de evaluación junto con la validez y la confiabilidad permite disponer de pruebas mejor diseñadas, más robustas, donde las perfeccionadas herramientas de medición facultan tomar decisiones en beneficio de los individuos y de la sociedad en su conjunto.

Referencias

American Educational Research Association, American Psychological Association, National Council on Measurement in Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing*. Washington D. C.: Autor.

American Psychological Association. (1954). *Technical recommendations for psychological test and diagnostic techniques*. Washington D. C.: Autor.

American Psychological Association. (1966). *Standards for Educational and Psychological Test and Manual*. Washington D. C.: Autor.

American Psychological Association. (2010). *Ethical Principles for Psychologists and Code of Conduct*. Washington D. C.: Autor.

André, N. y Loye, N. (2015). La validité psychologique: Un regard global sur le concept centenaire sa genése ses avatars. *Mesure et Évaluation en Éducation, 37*(3), 125-148. doi:10.7202/1036330ar

Argibay, J. (2006). Técnicas psicométricas: Cuestiones de validez y confiabilidad. *Subjetividad y Procesos Cognitivos, 8*, 15-33.

Binet, A. (1910). Qu´est-ce qu´une émotion? Qu´est-ce qu´un acte intellectuel? *L´Année Psycologique, 17*, 1-47.

Blanco-Villaseñor, Á. (1991). La teoría de la generalizabilidad aplicada a los diseños observacionales. *Revista Mexicana de Análisis de la Conducta*, *17*(3), 23-53.

Bond, T. y Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Borsboom, D., Mellenbergh, G. J. y Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071.

Bowen, W. G. y Bok, D. (1998). *The shape of the river.* Princeton, NJ: Princeton University Press.

Campbell, D. (1960). Recommendations for APA test standards regarding construct, trait or discriminate validity. *American Psychologist*, *15*(8), 546-553.

Carvajal-Carrascal, G. (2012). Medición de fenómenos de enfermería: El reto de la validez y la confiabilidad en la investigación cuantitativa. *Aquichan*, *12*(1).

Céspedes, V. (2009). *Modelo conceptual del manejo del síntoma: Clasificación por percepción, evaluación y respuesta de mujeres con síndrome coronario agudo; originada por la construcción de un instrumento validado en Bogotá, Colombia* (Tesis doctoral, Universidad Nacional de Colombia, Bogotá).

Chan, E. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. En B. Zumbo y E. Chan (Eds.), *Validity and validation in social, behavioral and health sciences* (pp. 9-24). Nueva York, NY: Springer.

Cupani, A. (2011). Acerca de la objetividad científica. *Scientiae Studia*, *9*(3), 501-525. doi:10.1590/S1678-31662011000300004

Educational Testing Service. (1987). *Standards for quality and fairness. Adopted by the Board of Trustees.* Princeton, NJ: Autor.

Embretson, S. (2007). Construct validity: A universal validity system or just another test evaluation procedure. *Educational Researcher*, *36*(8), 449-455. doi:10.3102/0013189X07311600

Fox, J. y Glas, C. (2001). Bayesian estimation of a multilevel in model using Gibbs sampling. *Psychometrika*, *66*(2), 271-288. doi:10.1007/BF02294839

Fox, J. y Glas, C. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*(2), 169-191. doi:10.1007/BF02294796

Froese-Germain, B. (1999). *Standardized testing: Undermining equity in education. Report prepared for the National Issues in Education Initiative.* Ottawa: Canadian Teachers' Federation.

García, M. (1955). Objetividad en el conocimiento científico. *Revista Cubana de Filosofía*, *3*(12), 21-26.

Gaukroger, S. (2012). *Objectivity. A very short introduction.* Oxford: Oxford University Press.

Gómez, J. e Hidalgo, M. (2003). Desarrollos recientes en Psicometría. *Avances en Medición*, *1*(1), 17-36.

Gómez, J. e Hidalgo, M. (2005). La validez de los test, escalas y cuestionarios. *La Sociología en sus Escenarios*, *12*, 1-14.

Gómez, J., Hidalgo, D. y Guilera, G. (2010). El sesgo de los instrumentos de medición. Test justos. *Papeles del Psicólogo*, *31*(1), 75-84.

Guilford, J. P. (1987). Validity of measurements. En J. P. Guilford (Ed.), *Fundamental statistics in psychology and education* (pp. 424-458). Tokyo: McGraw-Hill - Kogakusha.

Herrenstein, R. J. y Murray, G. (1994). *The Bell Curve. Intelligence and class structure in American life.* Nueva York, NY: Simon y Schuster.

International Test Commission. (2014). *International guidelines on the security of tests, examinations, and other assessments.* Recuperado de www.intestcom.org

International Test Commission. (2016). *The ITC guidelines for translating and adapting tests.* Recuperado de www.intestcom.org

Jeffrey, M. (2003). *Test validation: A literature review,* 1-46. Florida, CA: University of Florida.

Jiménez, K. y Montero, E. (2013). Aplicación del modelo de Rasch, en el análisis psicométrico de una prueba de diagnóstico en matemáticas. *Revista Digital Matemáticas, Educación e Internet, 13*(1), 1-24. doi:10.18845/rdmei.v13i1.1628

Jornet, J. y Suarez, R. (1996). Pruebas estandarizadas y evaluación del rendimiento: Usos y características métricas. *Revista de Investigación Educativa, 14*(2), 141-163.

Kane, M. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37*(2), 76-82. doi:10.3102/0013189X08315390

Kelley, T. L. (1927). Proposes served by educational test. En T. Kelley (Ed.), *Interpretation of educational measurements* (págs. 18-43). Nueva York, NY: World Book Company Yorkers on Hudson.

Kerlinger, F. y Howard, L. (2008a). Confiabilidad. En F. Kerlinger y L. Howard (Eds.), *Investigación del comportamiento. Métodos de investigación en ciencias sociales (*pp. 581-602). Ciudad de México: McGraw Hill.

Kerlinger, F. y Howard, L. (2008b). Validez. En F. Kerlinger y L. Howard (Eds.), *Investigación del comportamiento. Métodos de investigación en ciencias sociales* (pp. 603-628). Ciudad de México: McGraw Hill.

Kimberlin, C. y Winterstein, A. (2008). Validity and reliability of measurement instruments used in research. *American Journal Health-System Pharmacy, 65*(1), 2276-2284. doi:10.2146/ajhp070364

Kovač-Šebart, M. y Krek, J. (2009). *Justice in the assessment of knowledge: The opinions of teachers and parents.* Cracovia: AFM Publishing House.

Lane, R. (1999). *Validity evidence for assessments. Reidy interactive lecture series.* Pittsburgh, PA: University Pittsburgh.

Larroyo, F. (1968). *El positivismo lógico. Pro y contra.* Ciudad de México: Editorial Porrúa.

Linacre, J. y Wright, B. (1995). *How do Rasch and 3P differ? MESA Laboratory.* Chicago, IL: Kimbark.

Linacre, J. (1994*). Many-facet, Rasch measurement, MESA Press.* Chicago, IL: Kimbark.

Lissittz, R. y Samuelsen, K. (2007). Dialogue on validity. A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448. doi:10.3102/0013189X07311286

Messick, S. (1993, abril). *Foundation of validity: Meaning and consequences in psychological assessment.* Comunicación presentada en el Second Conference of the European Association of Psychological Assessment, Groningen, Netherlands.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 1-8.

Mislevy, R. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469. Doi: 10.3102/0013189X07311660

Morales de Barbenza, C. (2001). Consideraciones acerca de la objetividad en evaluación psicológica. *Interdisciplinaria*, *18*(2), 169-178.

Moss, P. (2007). Reconstructing validity. *Educational Researcher*, *36*(8), 470-476.

Muñiz, J. (2010). Las teorías de los test: Teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, *31*(1), 57-66.

Muñiz, L., Elosua, P. y Hambleton, R. (2013). Directrices para la traducción y adaptación de los test: Segunda edición. *Psicothema*, *25*(2), 151-157.

Myers, A. y Hansen, C. H. (2002). *Experimental psychology*. Belmont, CA: Wadsworth Thomson Learning.

Newton, P. y Baird, J. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 173-177. doi:10.1080/0969594X.2016.1172871

Nunnally, J. y Bernstein, I. (1995). *Teoría psicométrica*. Ciudad de México: McGraw-Hill.

Organización para la Cooperación y el Desarrollo Económicos. (2005). *PISA 2003 technical report*. París: OECD Publishing.

Organización para la Cooperación y el Desarrollo Económicos. (2009). *PISA 2006 technical report*. París: OECD Publishing.

Organización para la Cooperación y el Desarrollo Económicos. (2010). *PISA 2009 results: Learning to learn – Student engagement, strategies and practices (Volume III)*. París: OECD Publishing. doi:10.1787/9789264083943-en

Organización para la Cooperación y el Desarrollo Económicos. (2016). *Equations and inequalities: Making mathematics accessible to all*. París: OECD Publishing. doi:10.1787/9789264258495-en.

Organización para la Cooperación y el Desarrollo Económicos. (2016). *Low-performing students: Why they fall behind and how to help them succeed*. París: OECD Publishing. doi:10.1787/9789264250246-en.

Padilla, J., Gómez, J., Hidalgo, M. y Muñiz, J. (2006). La evaluación de las consecuencias del uso de los test en la teoría de la validez. *Psicothema*, *18*(2), 307-312.

Phelps, R. P. (2005). *Defending standardized testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Prieto, G. y Delgado, A. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, *15*(1), 94-100.

Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, *31*(1), 67-74.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

Real Academia Española. (2016). *Diccionario de la Lengua Española*. Recuperado de http://dle.rae.es/?id=QmvS5XH

Ritter, N. (February, 2010). *Understanding a widely misunderstood statistic: Cronbach´s alpha*. Comunicación presentada en el Annual Meeting of the Southwest Educational Research Association, Nueva Orleans. Recuperado de http://files.eric.ed.gov/fulltext/ED526237.pdf

Sánchez-Meca, J., López-Pina, J. y López, J. (2009). Generalización de la fiabilidad: Un enfoque meta analítico aplicado a la fiabilidad. *Fisioterapia*, *31*(6), 262-270.

Santiago, P., Halász, G., Levacíc, R. y Shewbridge, C. (2016) *Reviews of school resources: Slovak Republic*. París: OECD Publishing. doi:10.1787/9789264247567-en

Sattler, J. (2010). Niños de minorías étnicas. En J. Sattler (Ed.), *Evaluación infantil. Fundamentos cognitivos* (pp. 134-181). Ciudad de México: Manual Moderno.

Shavelson, R. y Webb, N. (2005). *Generalizability theory*. Newbury Park, CA: Sage Publications.

Shewbridge, C., Herczyński, J., Radinger, T. y Sonnemann, J. (2016). *OECD reviews of school resources: Czech Republic 2016*. París: OECD Publishing. doi:10.1787/9789264262379-en

Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, *9*, 167-194. doi:10.1080/15305050903106883

Sireci, S. (2007). On validity theory and test validation. *Educational Research*, *36*(8), 477-481. doi:10.1002/9781118445112.stat06403

Sireci, S. y Padilla, J. (2014). Validating assessment: Introduction to the special section. *Psicothema*, *26*(1), 97-99. doi:10.7334/psicothema2013.255

Tristán, A. (1996). *Nota 5: Contribución al estudio del error de medida. Kalt Criterial. Un programa de la familia Kalt. Versión 2. Guía de usuario*. San Luis Potosí: IEIA.

Tristán, A. (2010). *Theoretical Alpha values for objective test*. Recuperado de https://www.coreprojects.org/PROMIS/PROMIS2/Sandbox/Presentations/Tristan-AlphaPresentation.pdf

Wright, B. y Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement*, *1*(1), 84-106.

Wright, B. y Stone, M. (1999). *Measurement essential*. Wilmington, DE: Wide Range.

Zamora, E. (2007). *Evaluación objetiva de la calidad sensorial de los alimentos procesados*. La Habana: Editorial Universitaria.

Zumbo, B. (2009). Validity as contextualized and pragmatic explanation, and implication for validation practice. En R. Lissitz (Ed), *The concept of validity, revisions, new directions and applications* (pp. 65-82). Charlotte, NC: Information Age Publishing.

Zúñiga, M. y Montero, E. (2007). Teoría G. Un futuro paradigma para el análisis de pruebas psicométricas. *Anualidades en Psicología*, *21*(108), 117-144. doi:10.15517/ap.v21i108.29

Breve CV de los autores

Agustín Tristán López

Doctor en Ingeniería por la *École Nationale des Ponts et Chaussées*, París, Francia. Director General del Instituto de Evaluación e Ingeniería Avanzada, S.C. Asesor en psicometría y evaluación educativa, responsable de proyectos de certificación en docencia y en el área profesional para Colegios de Profesionales en las áreas de la Salud e Ingenierías. Su interés principal se centra en desarrollo de sistemas de medición, diseño de modelos matemáticos y estadísticos con teoría clásica y modelos logísticos y de Rasch. Autor de más de 30 productos de software para evaluación en educación y salud. Cuenta con más de 40 publicaciones en el tema de evaluación educativa. Email: atristan@ieia.com.mx. Sitio web: www.ieia.com.mx.

Nancy Yahibé Pedraza Corpus

Doctora en Estudios de Población, Centro de Estudios Demográficos, Urbanos y Ambientales, de El Colegio de México. Responsable Psicopedagógica del Instituto de Evaluación e Ingeniería Avanzada, S.C. Asesora y da seguimiento a diversos sistemas de evaluación educativos y procesos de certificación profesional. Especializada en diseño de pruebas e ítems objetivos para evaluar competencias con énfasis en aspectos sociales, actitudinales y psicológicos. Su interés principal se centra en el desarrollo de sistemas de medición, y en el análisis del comportamiento y la evaluación de sistemas educativos. Email: nancypedraza@ieia.com.mx. Sitio Web: www.ieia.com.mx.

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 33-49.
https://doi.org/10.15366/riee2017.10.1.002

The "Teaching to the Test" Family of Fallacies

La Familia de Falacias "Enseñando para el Examen"

Richard P. Phelps *

University of Pennsylvania

This article explains the various meanings and ambiguities of the phrase "teaching to the test" (TttT), describes its history and use as a pejorative, and outlines the policy implications of the popular, but fallacious, belief that "high stakes" testing induces TttT which, in turn, produces "test score inflation" or artificial test score gains. The history starts with the infamous "Lake Wobegon Effect" test score scandal in the US in the 1980s. John J. Cannell, a medical doctor, discovered that all US states administering national norm-referenced tests claimed their students' average scores exceeded the national average, a mathematical impossibility. Cannell blamed educator cheating and lax security for the test score inflation, but education insiders managed to convince many that high stakes was the cause, despite the fact that Cannell's tests had no stakes. Elevating the high stakes causes TttT, which causes test score inflation fallacy to dogma has served to divert attention from the endemic lax security with "internally administered" tests that should have encouraged policy makers to require more external controls in test administrations. The fallacy is partly responsible for promoting the ruinous practice of test preparation drilling on test format and administering practice tests as a substitute for genuine subject matter preparation. Finally, promoters of the fallacy have encouraged the practice of "auditing" allegedly untrustworthy high-stakes test score trends with score trends from allegedly trustworthy low-stakes tests, despite an abundance of evidence that low-stakes test scores are far less reliable, largely due to student disinterest.

Keywords: Test security, Educator cheating, Test score inflation, High stakes, Standardized tests, Education.

Este artículo explica los diversos significados y ambigüedades de la frase "enseñar para el examen" (*TttT: teaching to the test* en inglés), describe su historia y su uso como un peyorativo, y describe las implicaciones políticas de la creencia popular, pero falaz, que las pruebas de a "gran escala" inducen TttT que, a su vez, produce una "inflación en la calificación obtenida en el examen" o ganancias em cuanto a los puntos obtenidos en la prueba. La historia comienza con el infame escándalo de la puntuación de la prueba "Lake Wobegon Effect" en los Estados Unidos en los años ochenta. John J. Cannell, un médico, descubrió que todos los estados de los Estados Unidos que administraban pruebas nacionales con referencias normativas afirmaban que los puntajes promedio de sus estudiantes excedían el promedio nacional, una imposibilidad matemática. Cannell atribuyó a los educadores el engaño y la seguridad laxa por la inflación de la puntuación de los exámenes, pero los expertos en educación lograron convencer a muchos de que las pruebas a gran escala eran la causa, a pesar de que las pruebas de Cannell no tenían ninguna fiabilidad. Exagerar las pruebas a gran escala hace que TttT hace que la falla de la inflación de la puntuación de la prueba al dogma haya servido para desviar la atención de la seguridad laxa endémica con pruebas "internamente administradas" que deberían haber alentado a los responsables políticos a exigir más controles externos en las administraciones de las pruebas. La falacia es en parte responsable de promover la práctica ruinosa en la preparación de las pruebas en el formato de prueba y la administración de pruebas prácticas como un sustituto de la preparación de la materia original. Por último, los promotores de la falacia han fomentado la

*Contacto: richardpphelps@yahoo.com

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 1 de octubre de 2016
1ª Evaluación: 3 de enero de 2017
Aceptado: 21 de febrero de 2017

práctica de "auditar" tendencias de determinadas puntuación en las pruebas a gran escala con las tendencias de puntuación presuntamente confiables de las pruebas de baja exigencia, a pesar de la abundancia de pruebas donde las puntuaciones de las pruebas a menor escala son mucho menos confiables debido al desinterés de los estudiantes.

Palabras clave: Prueba de seguridad, Engaño de educador, Inflación de la puntuación del examen, Pruebas a gran escala, Pruebas estandarizadas, Educación.

1. Introduction

Standardized testing is one of the few means by which the public may ascertain what transpires inside school classrooms and, by far, the most objective.

For those inside education who would prefer to be left alone to operate schools as they wish, externally managed standardized tests intrude. Many actively encourage public skepticism of those tests' validity. Promoting the concept of "teaching to the test" as a pejorative is one part of the effort (Phelps, 2011c).

But, the meaning of the phrase is ambiguous (Shepard ,1990; Popham, 2004). At worst, it suggests grossly lax test security: teachers know the exact contents of an upcoming test and expose their students to that content, thereby undermining the test as an objective measure. Some testing critics would have the public believe that this is always possible. It is not. When tests are secure, the exact contents are unknown to teachers and test-takers alike until the moment scheduled testing begins and they hear instructions such as "please break open the seal of your test booklet".

More often, the phrase "teaching to the test" (TttT) is used pejoratively when it allegedly induces teachers to reduce the quality of instruction. There are two ways this can happen.

First, TttT allegedly lowers educational quality due to the limitations of tests. Critics suggest that tests—or, typically, externally managed standardized tests—are not well correlated with learning. These tests cannot measure all that students learn, perhaps not even most of, or the best parts of, what they learn. If true, then teaching only those components of learning that tests can capture neglects other, allegedly important, components of learning.

For a skeptic, the assertion begs the question: if tests do not measure important components of learning, how do we know those components exist? The philosopher and mathematician René Descartes is said to have written, "If a thing exists, it exists in some amount. If it exists in some amount, it is capable of being measured". Was he wrong? Are there types of learning that teacher-made tests can capture, but standardized tests cannot? …that teachers can ascertain, but tests cannot? Is some learning simply immeasurable?

Most outside education probably assume that if a student cannot demonstrate a certain knowledge or skill on a test, that student probably does not possess that knowledge or skill.

Some inside education argue that standardized tests can only assess "lower-order skills" or "factual recall". So, teachers avoid more enlightening and challenging instruction in favor of the mundane and simple. Without tests, they argue, teachers teach and students learn higher and deeper knowledge and skills that cannot be validly assessed by

standardized tests. Rather, better knowledge and skills can only validly be assessed by methods that require a large amount of teacher observation and judgment. Long-term or group projects are sometimes mentioned as good vehicles for the demonstration of "better" student knowledge and skills.

The second way that "teaching to the test" allegedly lowers instructional quality is through test preparation. "Test prep" occurs in a variety of forms. The simplest form familiarizes test-takers with the structure and format of the test, and is unrelated to subject matter content. Format familiarization is particularly important when the format of an upcoming test is, …well… unfamiliar. If students, for example, have never seen a multiple-choice test item before, some instruction and practice can be helpful.

Opinions differ about how much instruction and practice is appropriate. Most testing experts and test developers advocate only a brief amount of time. How much time does it take to understand how to respond to a multiple-choice test item, after all? When a test format is so convoluted that extensive training is required to use it, format decoding may have become the skill being tested. Psychometricians would then say the test has "construct-irrelevant variance"-that is, it is testing skills and knowledge different from the intended "construct" (i.e., the subject matter content).

Many testing opponents and some test preparation companies, however, argue that extended practice (i.e., "drilling") on test format and practice tests can improve test performance (Fraker, 1986-87; Smyth, 1990; Marte, 2011). Unfortunately, some school personnel believe them and convert their classrooms into "test prep factories", halting regular subject-matter instruction in favor of instruction on standardized test formats, drilling with test-maker-provided workbooks, or administering practice tests (Shepard 1990).

All educators consider this type of TttT unfortunate and debilitating to learning. Educators disagree, however, as to whether it works to increase test scores.

Teaching to the test's negative connotations can befuddle naïve education outsiders who assume a natural complementary relationship between teacher instruction and student testing. Shouldn't teachers teach subject matter that will be on the test? Shouldn't a test include subject matter a teacher covered in class? Why would a teacher teach "away from the test"—deliberating teach subject matter that will not be tested or, conversely, test subject matter that was not taught?

If a test is aligned with subject-matter standards, and its questions thoroughly cover them, can responsible teachers avoid teaching to the test? (Gardner, 2008).

2. A short history of US educators cheating on tests

Teaching to the test (TttT) is far more than a catch phrase or slogan, however. It has served for three decades to divert attention from a more serious problem in education in the United States-educators cheating on assessments used to judge their own performance. To elaborate adequately requires a short history lesson first.

Arguably, the current prevalence of large-scale testing in the United States began in the late 1970s. Some statistical indicators revealed a substantial decline in student achievement from the early 1960s on. Many blamed perceived permissiveness and lowered standards induced by the social movements of the 1960s and 1970s. Statewide

testing-at least of the most basic skills-was proposed to monitor the situation. For motivation, some states added consequences to the tests, typically requiring a certain score for high school graduation.

With few exceptions (e.g., California, Iowa, New York), however, states had little recent experience in developing or administering standardized tests or writing statewide content standards. That activity had been deferred to schools and school districts. So, they chose the expedient of purchasing "off the shelf" tests-nationally norm-referenced tests (NRTs)[1] (Phelps 2008/2009a; 2010). Outside the states of Iowa or California, the subject matter content of NRTs matched that of no state. Rather, each covered a pastiche of content, a generic set thought to be fairly common.

Starting in the 1970s, the state of Florida required its high school students to exceed a certain score on one of these. Those who did not were denied diplomas, even if they met all other graduation requirements.

A group of 10 African-American students who were denied high school diplomas after failing three times to pass Florida's graduation test sued the state superintendent of education (Buckendahl and Hunt, 2007). The plaintiffs claimed that they had had neither adequate nor equal opportunity to master the "curriculum" on which the test was based. Ultimately, four different federal courtrooms would host various phases of the trial of *Debra P. v. Turlington* between 1979 and 1984.

"Debra P." won the case after a study revealed a wide disparity between what was taught in classrooms to meet state curricular standards and the curriculum embedded in the test questions. A federal court ordered the state to stop denying diplomas for at least four years while a new cohort of students worked its way through a revised curriculum at Florida high schools and faced a test aligned to that curriculum.

The *Debra P.* decision disallowed the use of NRTs for consequential, or "high-stakes", decisions. But, many states continued to use them for other purposes. Some were still paying for them anyway under multi-year contracts. Typically, states continued to use NRTs as systemwide diagnostic and monitoring assessments, with no consequences tied to the results.

Enter a young medical resident working in a high-poverty region of rural West Virginia in the mid-1980s. He heard local school officials claim that their children scored above the national average on standardized tests. Skeptical, he investigated further and ultimately discovered that every U.S. state administering NRTs claimed to score above the national average, a statistical impossibility. The phenomenon was tagged the "Lake Wobegon Effect" after Garrison Keillor's "News from Lake Wobegon" radio comedy sketch, in which "all the children are above average".

The West Virginia doctor, John Jacob Cannell, M.D., would move on to practice his profession in New Mexico and, later, California, but not before documenting his investigations in two self-published books, *How All Fifty States Are above the National Average* and *How Public Educators Cheat on Standardized Achievement Tests.* (Cannell, 1987, 1989)

---

[1] Such as the Iowa Tests of Basic Skills (ITBS), Iowa Test of Educational Development (ITED), Stanford Achievement Test (the "other SAT"), or the California Test of Basic Skills (CTBS).

Cannell listed all the states and all the tests involved in his research. Naturally, all the tests involved were nationally normed, off-the-shelf, commercial tests, the type that the *Debra P. v. Turlington* decision had disallowed for use with student stakes. It is only because they were nationally normed that comparisons could be made between their jurisdictions' average scores and national averages.

By the time Cannell conducted his investigation in the mid- to late-1980s, about twenty states had developed *Debra P*-compliant high-stakes state tests, along with state content standards to which they were aligned. But, with the single exception of a Texas test[2], none of them was comparable to any other, nor to any national benchmark. They were "criterion-referenced" or "standards-based" tests unique to each state, and not nationally norm-referenced tests. And, again with Texas excepted, Cannell did not analyze them.

Dr. Cannell cited educator dishonesty and lax security in test administrations as the primary culprits of the Lake Wobegon Effect, also known as "test score inflation" or "artificial test score gains".

With stakes no longer attached, security protocols for the NRTs were considered unnecessary, and relaxed. It was common for states and school districts to have purchased the NRTs "off the shelf" and handle all aspects of test administration themselves. Moreover, to reduce costs, they could reuse the same test forms (and test items) year after year. Even if some educators did not intentionally cheat, over time they became familiar with the test forms and items and could easily prepare their students for them. With test scores rising over time, administrators and elected officials discovered that they could claim credit for increasing learning.

Conceivably, one could argue that the boastful education administrators were "incentivized" to inflate their students' academic achievement. But, incentives exist both as sticks and carrots. Stakes are sticks. There were no stakes attached to these tests. In many cases, the administrators were not even obligated to publicize the scores. Certainly, they were not required to issue boastful press releases attributing the apparent student achievement increases to their own managerial prowess. The incentive in the Lake Wobegon Effect scandal was a carrot-specifically, self-aggrandizement on the part of education officials.

Regardless the fact that no stakes attached to Cannell's tests, however, prominent education researchers blamed "high stakes" for the test-score inflation he found (Koretz et al., 1991). Cannell had exhorted the nation to pay attention to a serious problem of educator dishonesty and lax test security, but education insiders co-opted his discovery and turned it to their own advantage (Phelps, 2006).

"There are many reasons for the Lake Wobegon Effect, most of which are less sinister than those emphasized by Cannell" (Linn, 2000, p.7) said the co-director of a federally-

[2] The Texas TEAMS was a hybrid, partly a complete NRT, but with other test items added to thoroughly cover state content standards. The NRT portion was used to make national comparisons. But, only items aligned to state content standards were used to make consequential decisions.

funded research center on educational testing-for over three decades the *only* federally-funded research center on educational testing.[3]

Another of the center's scholars added:

> *Scores on high-stakes tests—tests that have serious consequences for students or teachers—often become severely inflated. That is, gains in scores on these tests are often far larger than true gains in students' learning. Worse, this inflation is highly variable and unpredictable, so one cannot tell which school's scores are inflated and which are legitimate.* (Koretz, 2008, p. 131)

These assertions supply the many educators predisposed to dislike high-stakes tests anyway a seemingly scientific (and seemingly not self-serving or ideological) argument for opposing them. Meanwhile, they present policymakers a conundrum: if scores on high-stakes tests improve, likely they are meaningless-leaving them no objective and reliable measure of school improvement. So they might just as well do nothing as bother doing anything.

After Dr. Cannell left the debate and went on to practice medicine, these education professors and their colleagues would repeat the mantra many times-high stakes (not lax security) cause test-score inflation-in dozens of reports published both by their center and by the National Research Council, whose educational testing research function they have co-opted (Baker, 2000; Linn, 2000; Linn, Graue, & Sanders, 1990; Shepard, 1990, 2000).

Cannell's main points-that educator cheating was rampant and test security inadequate-were dismissed out of hand and persistently ignored thereafter. The educational consensus fingered "teaching to the test" for the crime, manifestly under pressure from the high stakes of the tests.

Cannell's tests had no stakes. That's a fact anyone can verify. The tests he included in his analysis are listed in his reports. Indeed, with the *Debra P.* decision settled in the federal courts in the early 1980s, Cannell's tests could not legally have had stakes. Nonetheless, ask most anyone inside education today for the primary lesson to emerge from Dr. Cannell's famous "Lake Wobegon Effect" studies, and they will tell you: high-stakes induces teaching to the test, which induces test-score inflation-artificial increases in test scores unrelated to actual gains in student learning.

On the one hand, it is astonishing that they stick with the notion because it is so obviously wrong. The SAT and ACT university admission tests have stakes-one's score on either helps determine which university one attends. But, they have shown no evidence of test-score inflation. (Indeed, the SAT was re-centered in the 1990s because of score *deflation.*) The most high-stakes tests of all-occupational licensure tests-show no evidence of test-score inflation. Both licensure tests and the SAT and ACT, however, have been administered with tight security and ample test form and item rotation.

---

[3] Since the early 1980s, the Center for Research on Educational Standards and Student Testing (CRESST) has been continually headquartered in UCLA's education school, and continually partnered with the University of Colorado's and the University of Pittsburgh's education schools. Other partners have included the Rand Corporation, and the education schools at Arizona State University, Stanford University, and at other University of California campuses.

3. Spot the Causal Factor

Table 1. Security and Stakes in evaluation

	HIGH SECURITY (EXTERNAL ADMINISTRATION)	LAX SECURITY (INTERNAL ADMINISTRATION)
High stakes	No test-score inflation e.g., SAT, ACT, licensure exams	Test-score inflation possible e.g., some internally administered district and state exams
No/low stakes	No test-score inflation e.g., National Assessment of Educational Progress (NAEP)	Test-score inflation possible e.g., Cannell's "Lake Wobegon" exams

Source: Auhor.

On the other hand, this "folk belief" is not unlike others in the US education school catechism, such as learning styles, multiple intelligences, and discovery learning: consistently proven wrong, but persisting nonetheless and matching the radical egalitarian and progressive education ideals that have consumed US schools of education.

The belief fits well into the knowledge base that many US education professors *want to* believe is true, rather than that which is true. US educationist doctrine may be less about a search for truth, and more an aspiration to what *should be* true -a set of knowledge they consider better because they consider it morally superior.

The late senator from New York, Daniel Patrick Moynihan, famously said "Everyone is entitled to their own opinion, but not their own set of facts"[4]. Apparently, US education professors do not agree. They have successfully elevated panoply of falsehoods aligned with their preferences to "facts" in the collective working memory. Their faux facts may influence US education policy-making more than real ones.

The scholars at the federally funded research center followed Cannell's studies with two of their own purporting to demonstrate both that teaching to the test works to artificially inflate test scores, and that high stakes induce teaching to the test. Both studies are methodologically flawed beyond the point of salvaging (Phelps, 2008, 2009a, 2010). Nevertheless, they remain, along with the distortion of Dr. Cannell's studies, highly respected among the US education professoriate and the foundation for most educators' understanding of the nature and implications of teaching-to-the-test (Crocker, 2005).

The reasoning goes like this: under pressure to raise test scores by any means possible, teachers reduce the amount of time devoted to regular instruction and, instead, focus on test preparation that can be subject-matter free (i.e., test preparation or test coaching). Test scores rise, but students learn less (Koretz, 1992, 1996; Koretz et al., 1991).

The two foundational studies examined certain patterns in the pre- and post-test scores from the first decade (i.e., late 1970s and early 1980s) of the federal government's compensatory education program (Linn, 2000) and the "preliminary findings" from the

[4] http://www.goodreads.com/author/quotes/219349.Daniel_Patrick_Moynihan

early 1990s of a test "perceived to be high stakes" in one school district (Koretz, Linn, Dunbar, & Shepard, 1991).

Research conducted on this hypothesis by others concludes that teachers who spend more than a brief amount of time focused on test preparation do their students more harm than good[5]. Their students score lower on the tests than do other students whose teachers eschew any test preparation beyond simple format familiarization and, instead, use the time for regular subject-matter instruction (see, for example, Allensworth, Correa, & Ponisciak, 2008; Camara, 2008; Crocker, 2005; Moore, 1991; Palmer, 2002). Moreover, students who know the specific content of prep tests beforehand may be lulled into a false confidence, study less, learn less, and score lower on final exams than those who do not (see, for example, Tuckman, 1994; Tuckman & Trimble, 1997).

The more widespread the belief that tests can be gamed by learning tricks unrelated to subject matter acquisition, the more customers and profits they gain.

As it turns out, neither of the two foundational studies of high-stakes testing effects included high-stakes tests. The researchers crossed their fingers behind their backs and employed an archaic, overly broad definition for the term "high stakes" for which virtually any standardized test would qualify (Phelps, 2010).[6] Yes, what they used was a definition, but it was neither the standard industry definition nor one that anyone outside their circle would reasonably assume for the term.[7]

This "floating definition" semantic sleight-of-hand is commonplace in US education research, its frequency of use grossly underappreciated by journalists and policy-makers. Education researchers surreptitiously substitute an obscure connotation for a term that varies from the more commonly understood denotation and explain the substitution, when they explain it at all, only in the fine print (Phelps, 2010).

One of the two studies was conducted in a school district and with tests that remain unidentified (Koretz, 2008). To this day, the researchers claim that they must keep that information secret to "protect" their sources (from what is not explained) (Staradamskis, 2008).

---

[5] Messick & Jungeblut (1981); DerSimonian & Laird (1983); Kulik, Bangert-Drowns, & Kulik (1984); Whitla (1988); Snedecor (1989); Becker (1990); Powers (1993); Allalouf & Ben-Shakhar (1998); Camara (1999); Powers & Rock (1999); Robb & Ercanbrack (1999); Briggs (2001); Zehr (2001); Briggs & Hansen (2004); Wainer (2011); Marte (2011); and Arendasy, Sommer, Gutierrez-Lobos, & Punter (2016).

[6] CRESST researchers cited (Shepard, 1990, p.17) a definition they attribute to James Popham from 1987 ascribing "high stakes" to any test whose aggregate results were reported publicly or which received media coverage. With the widespread passage of "truth in testing" and other open records laws, starting with California and New York State in the late 1970s, the aggregate results of all large-scale tests became public record. By their out-of-date definition, ALL large-scale tests are "high stakes".

[7] The standard, industry-wide definition of "high stakes" could be found in the *Standards for Educational and Psychological Testing* (AERA et al.), "High-stakes test. A test used to provide results that have important, direct consequences for examinees, programs, or institutions involved in the testing" (p.176) "Low-stakes test. A test used to provide results that have only minor or indirect consequences for examinees, programs, or institutions involved in the testing" (p.178).

Secret definitions. Secret locations. Secret tests. Such studies may stand forever because they are neither replicable nor falsifiable. More like religion than science; they require faith. And, inside U.S. education one finds many willing believers.

Meanwhile, a cornucopia of studies contradicting the two research center studies have been repeatedly declared nonexistent by the same researchers and thousands of sympathetic others inside US education schools (Phelps, 2005, 2008, 2009b, 2012a, 2012b).

Elevating teaching-to-the-test to dogma, from the beginning with the distortion of Dr. Cannell's findings, has served to divert attention from scandals that should have threatened US educators' almost complete control of their own evaluation.[8] Had the scandal Dr. Cannell uncovered been portrayed honestly to the public-educators cheat on tests administered internally with lax security-the obvious solution would have been to externally manage all assessments (Oliphant, 2011).

More recent test cheating scandals in Atlanta, Washington, D.C., and elsewhere once again drew attention to a serious problem. But, instead of blaming lax security and internally managed test administration, most educators blamed the stakes and alleged undue pressure that allegedly ensues (Phelps, 2011a). Their recommendation, as usual: drop the stakes and reduce the amount of testing. Never mind the ironies: they want oversight lifted so they may operate with none, and they admit that they cannot be trusted to administer tests to our children properly, but we should trust them to educate our children properly if we leave them alone.

Perhaps the most profound factoids revealed by the more recent scandals were, first, that the cheating had continued for ten years in Atlanta before any responsible person attempted to stop it and, even then, it required authorities outside the education industry to report the situation honestly. Second, in both Atlanta and Washington, DC, education industry test security consultants repeatedly declared the systems free of wrongdoing (Phelps, 2011b).

Meanwhile, thirty years after J. J. Cannell first showed us how lax security leads to corrupted test scores, regardless the stakes, test security remains cavalierly loose in the United States. We have teachers administering state tests in their own classrooms to their own students, principals distributing and collecting test forms in their own schools. Security may be high outside the schoolhouse door, but inside, too much is left

[8] More than in most countries, the U.S. public education system is independent, self-contained, and self-renewing. Education professionals staffing school districts make the hiring, purchasing, and school catchment-area boundary-line decisions. School district boundaries often differ from those of other governmental jurisdictions, confusing the electorate. In many jurisdictions, school officials set the dates for votes on bond issues or school board elections, and can do so to their advantage. Those school officials are trained, and socialized, in graduate schools of education. A half-century ago, most faculties in graduate schools of education may have received their own professional training in core disciplines, such as Psychology, Sociology, or Business Management. Today, most education school faculty are themselves education school graduates, socialized in the prevailing culture. The dominant expertise in schools of education can maintain its dominance by hiring faculty who agree with it and denying tenure to those who stray. The dominant expertise in education journals can control education knowledge by accepting article submissions with agreeable results and rejecting those without. Even most testing and measurement PhD training programs now reside in education schools, inside the same cultural cocoon.

to chance. And, as it turns out, educators are as human as the rest of us; some of them cheat and not all of them manage to keep test materials secure, even when they aren't intentionally cheating.

4. Codifying TttT falsehoods

The primary advocates of the high-stakes-causes-TttT-which-causes-test-score-inflation belief (hereafter HS->TttT->TSI), reside at the Center for Research on Educational Standards and Student Testing (CRESST), for over three decades the only federally funded research center on educational testing. CRESST staff led the effort to discredit the work and findings of J. J. Cannell, the earnest medical doctor who uncovered the Lake Wobegon Effect scandal in the 1980s.

First, they rejected out of hand Cannell's contentions that educator cheating on tests was rampant and test security too lax. Second, in promoting HS->TttT->TSI, they instilled doubt in the reliability and validity of high-stakes test results.

Rather than stop there, however, they have advocated for thirty years that allegedly unreliable high-stakes test results should be "audited" by parallel low- or no-stakes tests. They reasoned that no-stakes test scores are reliable because there exist no incentives to cheat on them.

A cornucopia of research exists contradicting CRESST's faith in the reliability of low- and no-stakes test scores.[9] No matter, CRESST researchers have simply ignored it.

In summary, they promote all of the following beliefs:

1) HS->TttT->TSI. Again, as the theory's primary advocate writes,
2) "Scores on high-stakes tests-tests that have serious consequences for students or teachers-often become severely inflated. That is, gains in scores on these tests are often far larger than true gains in students' learning. Worse, this inflation is highly variable and unpredictable, so one cannot tell which school's scores are inflated and which are legitimate".
3) Subject-matter independent training in test taking works to increase test scores (as some test prep companies also claim).
4) High-stakes test scores are, at best, only partly related to subject matter mastery, because they are also highly correlated with subject-matter-free test-taking skills.
5) The cause of educator cheating in testing administrations is high-stakes; without high-stakes, educators do not cheat.
6) No- or low-stakes tests, by contrast, are not susceptible to test-score inflation because there are no incentives to manipulate scores.

The public policy implications of these beliefs are substantial. Given the statements above, responsible public policy should incorporate the following:

---

[9] See, for example, Brown & Walberg (1993); Wise & DeMars (2005); Eklof (2007); Abdelfattah (2010); Barry, Horst, Finney, Brown, & Kopp (2010); Wise & DeMars (2010); Wainer (2011); Zilberberg, Anderson, Finney, & Marsh (2013); Steedle (2014); Liu, Rios, & Borden (2015); Sessoms & Finney (2015); Smith, Given, Julien, Ouellette, & DeLong (2015); Mathers, Finney, & Myers (2016); and Rios, Guo, Mao, & Liu (2016).

a) In the interest of improving test scores, teachers should teach to high-stakes tests—that is, drill on test format. They should reduce the amount of time devoted to subject matter mastery-to regular instruction and learning-and, instead, devote more time to taking practice tests, coaching students on test-taking strategies, familiarizing their students with standardized test formats, etc.

b) Use of test prep services should be encouraged. Moreover, in the interest of fairness, these services should be subsidized, at least for poorer students.

c) If score trends for high-stakes tests are unreliable and those for no- or low-stakes tests are reliable, no- or low-stakes tests may be used validly as shadow tests to audit the reliability of high-stakes tests' score trends.

d) Test security (or, the integrity of test materials) is not an issue with no- or low-stakes tests, so they can be validly administered without security controls.

e) Or, eliminate the use of high stakes tests entirely. Given that they provide neither valid nor reliable information, there is no excuse for using them. Currently, high stakes tests are used for certification and licensure in most professions and trades.

Several years ago, CRESST staff occupied prominent positions on the committee drafting an update to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), arguably the most important document in testing. Left in charge of drafting a new chapter 13 on public policy, CRESST staff incorporated their set of beliefs and the logical policy recommendations derived therefrom (Phelps, 2011b).

Judging from comments on the draft Standards made publicly available, I was the only public critic of the CRESST draft chapter 13. I recommended deleting it completely. As it turns out, an intervention occurred and the chapter was overhauled to remove the most egregious pseudoscientific claims and recommendations (AERA, NCME, & APA, 2013).

But, what if I hadn't raised a fuss? Did I represent the only barrier between the *Standards* incorporating CRESST's TttT Family of Fallacies and *Standards* based on genuine research evidence? That would be frightening. But, I witnessed no one else raising anything more than trivial objections to draft chapter 13.

A Pyrrhic victory? Meanwhile, the TttT Family of Fallacies has received warm welcomes at the Organisation for Economic Co-operation and Development (OECD) and the educational testing office at the World Bank (Phelps, 2014). These international organizations promote these falsehoods worldwide.

5. What if lax test security causes test score inflation?

Thousands of externally imposed high-stakes tests show no evidence of test-score inflation. Likewise, low- and no-stakes tests notoriously lead to test-score inflation when test security (or, "the integrity of test materials") is lax. The necessary and sufficient condition for test-score inflation is lax security, not high stakes.

Reject the pseudoscience of the TttT Family of Fallacies and quite different public policy implications emerge. Following where the research evidence points:

1) Test scores and test score trends should not be trusted in the absence of test security controls, no matter what the stakes.
2) High-stakes test scores and score trends are typically not only valid and reliable when administered with tight security, they are more likely to be valid and reliable because they are more likely to be administered with tight security than low- and no-stakes tests
3) Educators are normal human beings, and respond to a variety of incentives, just like the rest of us. By cheating on no- or low-stakes tests, educators might then publicize and take credit for the ostensible student learning increases. Note, however, that no "stakes" are involved; rather, self-aggrandizement is the motive.
4) Drilling on test format not only does not improve learning, because it takes time away from subject matter instruction, it reduces it.
5) Money spent on test preparation services is money wasted if the service consists primarily of test-taking strategies, format familiarity, and practice test taking.

Given the statements above, responsible public policy should incorporate the following:

a) Consider test security (or, the "integrity of test materials") far more seriously than it has been, and applicable to many no-or low-stakes tests.
b) Encourage teachers to devote only a modicum of time to familiarizing their students with standardized test-taking formats and strategies. They should not sacrifice instruction in subject-matter mastery.
c) Eliminate the fallacious research practice that considers no-stakes tests to be always valid and reliable and thus trustworthy to use in "auditing" high-stakes tests.

References

Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality, 38,* 159-168.

Allalouf A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35*(1), 31-47.

Allensworth, E., Correa, M., & Ponisciak, S. (2008). *From high school to the future: ACT preparation—Too much, too late: Why ACT scores are low in Chicago and what it means for schools.* Chicago, IL: Consortium on Chicago School Research at the University of Chicago.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2013). *Standards for educational and psychological testing.* Washington, DC: AERA.

Arendasy, M. E., Sommer, M., Gutierrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence, 55,* 44-56.

Baker, E. L. (2000). *Understanding educational quality: Where validity meets technology*. Princeton, NJ: Educational Testing Service, Policy Information Center.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342-363. doi:10.1080/15305058.2010.508569

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research*, *60*(3), 373-417.

Briggs, D. C. (2001). The effect of admissions test preparation. *Chance*, *14*(1),10-18.

Briggs, D., & Hansen, B. (2004). *Evaluating SAT test preparation: Gains, effects, and self-selection*. Princeton, NJ: Educational Testing Service.

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research*, *86*(3), 133-136.

Buckendahl, C. W., & Hunt, R. (2005). Whose rules? The relation between the "rules" and "law" of testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 147-158). Mahwah, NJ: Psychology Press.

Camara, W. (1999). *Is commercial coaching for the SAT I worth the money?*. New York, NY: College Counseling Connections.

Camara, W. J. (2008). College admission testing: Myths and realities in an age of admissions hype. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 45-76). Washington, DC: American Psychological Association.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools. How all fifty states are above the national average*. Daniels, WV: Friends for Education.

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends for Education.

Crocker, L. (2005). Teaching for the test: How and why test preparation is appropriate. In R. P. Phelps (Ed.), *Defending standardized testing (*pp. 159-174). Mahwah, NJ: Psychology Press.

DerSimonian, R., & Laird, M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review*, *53*, 1-5.

Eklof, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*, 311-326. doi:10.1080/15305050701438074

Fraker, G. A. (1987). *The Princeton Review reviewed. The Newsletter*. Deerfield, MA: Deerfield Academy.

Gardner, W. (2008). *Good teachers teach to the test: That's because it's eminently sound pedagogy*. Retrieved from http://www.csmonitor.com/Commentary/Opinion/2008/0417/p09s02-coop.html

Koretz, D. (April, 1992). NAEP and the movement toward national testing. Paper presented at the *Annual Meeting of the American Educational Research Association*, San Francisco.

Koretz, D. M. (1996). *Improving America's schools: The role of incentives*. Washington, DC: National Academy Press.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (April, 1991). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C-L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*, 179-188.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4-16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that everyone is above average. *Educational Measurement: Issues and Practice*, *9*(3), 5-14.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79-94. doi: 10.1080/10627197.2015.1028618

Marte, J. (2011). *10 things test-prep services won't tell you. Market watch*. Retrieved from http://www.marketwatch.com/story/10-things-testprep-services-wont-tell-you-1301943701454

Mathers, C., Finney, S., & Myers, A. (2016, July). *How test instructions impact motivation and anxiety in low-stakes settings*. Paper presented at the Annual Meeting of the Psychometric Society, Asheville, NC.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, *89*, 191-216.

Moore, W. P. (1991). *Relationships among teacher test performance pressures, perceived testing benefits, test preparation strategies, and student test performance* (PhD dissertation). University of Kansas, Lawrence.

Oliphant, R. (2011). Modern metrology and the revision of our Standards for Educational and Psychological Testing: An open letter to American parents. *Nonpartisan Education Review / Essays*, *7*(4). Retrieved from http://www.nonpartisaneducation.org/Review/Essays/v7n4.pdf

Palmer, J. S. (2002). *Performance incentives, teachers, and students: Estimating the effects of rewards policies on classroom practices and student performance* (PhD dissertation). Ohio State University, Columbus, Ohio.

Phelps, R. P. (2005). The rich, robust research literature on testing's achievement benefits. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 55-90). Mahwah, NJ: Psychology Press.

Phelps, R. P. (2006). A tribute to John J. Cannell, M.D. *Nonpartisan Education Review/Essays*, *2*(4). Retrieved from http://www.nonpartisaneducation.org/Review/Essays/v2n4.pdf

Phelps, R. P. (2008/2009a). The rocky score-line of Lake Wobegon. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp.102-134). Washington D. C.: American Psychological Association.

Phelps, R. P. (2008/2009b). Educational achievement testing: Critiques and rebuttals. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 66-90). Washington D. C.: American Psychological Association.

Phelps, R. P. (2010). The source of Lake Wobegon. *Nonpartisan Education Review/Articles, 6(3)*. Retrieved from http://nonpartisaneducation.org/Review/Articles/v6n3.htm

Phelps, R. P. (2011a). *Standards for Educational & Psychological Testing*. New Orleans, LA: American Psychological Association.

Phelps, R. P. (2011b). Educator cheating is nothing new; doing something about it would be. *Nonpartisan Education Review/Essays, 7*(5). Retrieved from http://nonpartisaneducation.org/Review/Essays/v7n5.htm

Phelps, R. P. (2011c). *Teach to the test? The Wilson Quarterly*. Retrieved from http://wilsonquarterly.com/quarterly/fall-2013-americas-schools-4-big-questions/teach-to-the-test/

Phelps, R. P. (2012a). Dismissive reviews: Academe's Memory Hole. *Academic Questions, 25*(2), 228-241.

Phelps, R. P. (2012b). The rot festers: Another National Research Council report on testing. *New Educational Foundations, 1*(1). Retrieved from http://www.newfoundations.com/NEFpubs/NewEduFdnsv1n1Announce.html

Phelps, R. P. (2014). Synergies for better learning: An international perspective on evaluation and assessment. *Assessment in Education: Principles, Policies, & Practices, 21*(4), 481-493. doi:10.1080/0969594X.2014.921091

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*, 675-682.

Popham, W. J. (2004). All about accountability / "Teaching to the test": An expression to eliminate. *Educational Leadership, 62*(3), 82-83.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice, 39*, 24-30.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*(2), 93-118.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2016). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 16*, 1-36. doi:10.1080/15305058.2016.1231193

Robb, T. N., & Ercanbrack, J. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *Teaching English as a Second or Foreign Language, 3*(4).

Sessoms, J., & Finney, S. J. (2015) Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15*(4), 356-388. doi:10.1080/15305058.2015.1034866

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, *9*(3), 15-22.

Shepard, L. A. (April, 2000). The role of assessment in a learning culture. Presidential Address presented at the *Annual Meeting of the American Educational Research Association*, New Orleans.

Smith, J. K., Given, L. M., Julien, H., Ouellette, D., & DeLong, K. (2013). Information literacy proficiency: Assessing the gap in high school students' readiness for undergraduate academic work. *Library & Information Science Research*, *35*, 88-96.

Smyth, F. L. (1990). SAT coaching: What really happens to scores and how we are led to expect more. *The Journal of College Admissions*, *129*, 7-16.

Snedecor, P. J. (1989). Coaching: Does it pay-revisited. *The Journal of College Admissions*, *125*, 15-18.

Staradamskis, P. (2008, Fall). Measuring up: What educational testing really tells us. Book review. *Educational Horizons*, *87*(1). Retrieved from http://nonpartisaneducation.org/Foundation/KoretzReview.htm

Steedle, J. T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education*, *27*, 58-76. doi: 10.1080/08957347.2013.853072

Tuckman, B. W. (April, 1994). Comparing incentive motivation to metacognitive strategy in its effect on achievement. Paper presented at the *Annual Meeting of the American Educational Research Association*, New Orleans.

Tuckman, B. W., & Trimble, S. (August, 1997). Using tests as a performance incentive to motivate eighth-graders to study. Paper presented at the *Annual Meeting of the American Psychological Association*, Chicago.

Wainer, H. (2011). *Uneducated guesses: Using evidence to uncover misguided education policies.* Princeton, NJ: Princeton University Press.

Whitla, D. K. (1988). Coaching: Does it pay? Not for Harvard students. *The College Board Review*, *148*, 32-35.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1-17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27-41. doi:10.1080/10627191003673216

Zehr, M. A. (2001). *Study: Test-preparation courses raise scores only slightly.* New York, NY: Education Week.

Zilberberg, A., Anderson, R. D., Finney, S. J., & Marsh, K. R. (2013). American college students' attitudes toward institutional accountability testing: Developing measures. *Educational Assessment*, *18*, 208-234. doi:10.1080/10627197.2013.817153

Breve CV del autor

Richard P. Phelps

Founder of the Nonpartisan Education Group and editor of its peer-reviewed journal, the *Nonpartisan Education Review* (http://nonpartisaneducation.org), a Fulbright Scholar, and a fellow of the Psychophysics Laboratory. He has authored, or edited and authored, four books on assessment policy –*Correcting Fallacies about Educational and Psychological Testing* (APA); *Standardized Testing Primer* (Peter Lang); *Defending Standardized Testing* (Psychology Press); and *Kill the Messenger: The War on Standardized Testing* (Transaction)– and several statistical compendia. Phelps has held positions with several organizations working in assessment, including ACT, AIR, ETS, the OECD, Pearson, and Westat. He holds degrees from Washington, Indiana, and Harvard Universities, and a PhD in Public Policy from the University of Pennsylvania's Wharton School. ORCID ID: 0000-0003-4008-087x. Email: richardpphelps@yahoo.com

Revista Iberoamericana de Evaluación Educativa, 2017, *10*(1), 51-67.
https://doi.org/10.15366/riee2017.10.1.003

Evaluación y Pruebas Estandarizadas: Una Reflexión sobre el Sentido, Utilidad y Efectos de estas Pruebas en el Campo Educativo

Evaluation and Standardized Tests: A Reflection about the Meaning, Utility and Effects of these Tests in Educational Field

Manuel Fernández Navas *[1]
Noelia Alcaraz Salarirche [2]
Miguel Sola Fernández [2]

[1]Universidad de Cádiz y [2]Universidad de Málaga

La estandarización de pruebas para valorar los sistemas educativos ha ido in-crescendo desde que el discurso de las competencias cobrara protagonismo en la Unión Europea. El análisis de la ideología que sustentan dichas prácticas, así como del paradigma que se esconde tras ellas, son objetivos de este artículo. Para ello se realiza una reflexión por una de las máximas representantes de esta corriente en Europa, las denominadas pruebas PISA (*Programme for International Student Assessment*) (Programa Internacional para la Evaluación de Estudiantes). Los autores plantean una serie de argumentos que tratan de explicar por qué las pruebas estandarizadas representan un modelo de eficiencia en educación y no tanto, de calidad, donde la evaluación es entendida como un proceso de recogida de información útil y valiosa para emprender procesos de mejora. Y analizan algunas de las cuestiones más significativas que se desprenden del uso de pruebas estandarizadas en educación, tales como el uso político que se hace de los resultados, el sobre valor que se le otorga a las materias instrumentales o el desmantelamiento que se produce de la función social de la escuela, entre otras.

Palabras Clave: Pruebas estandarizadas, Evaluación, Aprendizaje, Competencias, Sistema educativo.

The standardization of tests to value the educational system has been in- crescendo since the competences speech becomes relevant in the European Union. The analysis of ideology supporting these practices, as well as the paradigm is behind them, is the objective of this paper. This requires a brief review of one of the more representative tests of this trend in Europe, the so-called PISA (Programme for International Student Assessment). The authors present arguments trying to explain why standardized tests represent an efficiency model based in education and not quality model based, where evaluation is understood as a process of collecting useful and valuable information that can undertake processes improvement. In the same way, authors analyse main questions about use of standardized tests in education like politic use of results, primacy of some areas about others or detriment of the school social function.

Keywords: Standarized tests, Evaluation, Learning, Competences, Educational system.

*Contacto: manuel.navas@uca.es

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 26 de octubre de 2016
1ª Evaluación: 2 de diciembre de 2016
Aceptado: 22 de diciembre de 2016

1. Introducción

En España, y en la UE, las pruebas estandarizadas como elemento para valorar el aprendizaje se desarrollan a partir de la aparición del término competencia, impulsado, recordemos, por la OCDE[1]. Aunque es preciso señalar que esto no es rigurosamente cierto. El término "competencia" aplicado a la formación del profesorado es más bien antiguo, constituyó en su momento una propuesta que tuvo ciertamente difusión y alcance incluso en España y ha sido analizado profusamente. Zeichner (1993) identifica a la C/PBTE –*Competency/Performance based teacher education:* formación del profesorado basada en competencias– como el máximo exponente del enfoque de eficacia social (denominado también conductista o tecnológico), cuya pretensión es fundamentalmente identificar y enseñar a los futuros profesores las competencias propias del profesor que tiene éxito en su enseñanza. El modelo de formación C/PBTE tiene un elevado componente práctico que trata de asegurar el dominio de las destrezas por parte de los alumnos. Ello derivó en técnicas formativas tales como la microenseñanza y la supervisión (Villar, 1977).

Tal y como plantea Fernández Navas (2016):

> *Salvando las distancias, la enseñanza basada en competencias de la actualidad comparte las pretensiones del aprendizaje colaborativo del modelo antiguo, ahora trasladadas a la formación de los ciudadanos para el mundo actual, y también las probabilidades de que derive en estrategias formativas de entrenamiento. Por otra parte, la propuesta de enseñanza-aprendizaje basada en competencias es relativamente novedosa en la Unión Europea, pero desde luego no en USA ni en los países más cercanos a su influencia (precisamente por razones de relaciones comerciales y de empleabilidad). Son numerosos los países centroamericanos que ya tenían supuestamente un sistema escolar orientado a la adquisición de competencias y, desde luego, sistemas de selección –exámenes nacionales, pruebas externas estandarizadas– que pretendían medir la consecución de tales competencias por parte del alumnado. Un claro ejemplo de ello son las denominadas Pruebas PAES, ECAP y de Evaluación de Logros en El Salvador, desde al menos 2001, tal y como ponen de manifiesto Martínez & Sola (2004).* (p. 273)

No obstante, como decimos, es en los inicios de este siglo XXI cuando han alcanzado su máximo esplendor en el ámbito educativo en España así como en aquellos que están bajo la influencia de la OCDE.

> *Las bases de la teoría de la medición fueron puestas desde el siglo XVIII por Laplace y Gauss. La aplicación de estas ideas al campo educativo comenzó desde el XIX, distinguiéndose países como Alemania, Inglaterra, los Estados Unidos y, en menor medida, Francia y las regiones francófonas de Suiza y Bélgica. A partir de los primeros años del siglo XX se desarrolla la metodología que se conoce ahora como teoría clásica de las pruebas (classical tests theory), a partir de la teoría de la confiabilidad y el modelo estadístico de las puntuaciones, con las nociones de puntaje verdadero, error de medida y confiabilidad de la prueba.* (Martínez Rizo, 2001, p. 54)

Dado el momento histórico en el que se asientan las bases de las pruebas estandarizadas, no resulta difícil identificar el paradigma educativo sobre el que se sustentan y la ideología que subyace a ellas. Como expone Pérez Gómez (1992 cit. por Alcaraz, 2015), para entender la virtualidad de los fenómenos de aprendizaje en educación es necesario

[1] Organización para la Cooperación y Desarrollo Económico.

reconocer los principios de las diferentes corrientes psicológicas y teorías del aprendizaje.

Nos encontramos ante apreciaciones puramente mecanicistas del aprendizaje propias de las teorías del condicionamiento que defienden una educación convertida en mera tecnología que trata de programar refuerzos en el momento oportuno (Pérez Gómez, 1992). Donde:

> *El acto educativo es el escenario en el que se programan con la máxima exactitud posible los estímulos externos, refuerzos, entradas (inputs), con la pretensión de obtener, también con la máxima precisión, las respuestas deseadas, salidas (outputs). El aprendizaje no es más que modificación de conductas, quedando el comportamiento humano reducido a la más simple de las explicaciones psicológicas.* (Alcaraz, 2014, p. 59)

Con lo cual todo acto de aprendizaje queda reducido a la relación estímulo-respuesta que hace la persona que aprende y al docente se le presupone la capacidad de controlar y predecir el comportamiento de dicho sujeto.

Es bajo esta ideología determinada desde donde se aporta la idea de que valorar el aprendizaje del alumnado a través de pruebas "masivas", estandarizadas, supone una herramienta valiosa en tanto en cuanto aporta información sobre los resultados que está produciendo un sistema educativo determinado.

Este auge de las pruebas estandarizadas asociado a la aparición del termino competencias no deja de resultar llamativo para algunos autores, que sostienen la tesis de que esta asociación no es aleatoria ni casual, sino que tiene que ver con un intento de legitimar la educación y sus sistemas de acreditación académica en la nueva sociedad (fundamentalmente tras la aparición de internet) en la que nos encontramos inmersos, y que Castells (2001) llama sociedad red, o a la que Bauman (2007) se refiere como Modernidad líquida, en la que las instituciones educativas ya no son, como antaño, las depositarias del saber y del conocimiento (Fernández Navas, 2016).

El problema de estas pruebas, como veremos en los apartados siguientes de este trabajo, radica en que ni pertenecen propiamente al concepto de evaluación educativa, ni es posible medir con precisión algo tan complejo como una competencia.

2. PISA y el paradigma de la eficiencia

Tal y como dejábamos ver en la introducción de este trabajo, el supuesto valor de estas pruebas estandarizadas se atribuye a que supuestamente "valoran" el aprendizaje del alumnado de forma masiva y lo atribuyen a los efectos del sistema educativo al que pertenecen los estudiantes. Permitiendo así, supuestamente, arrojar luz sobre sus carencias o bondades. El esquema es tan sencillo como lineal: el sistema educativo produce aprendizajes en los estudiantes y la medición de tales resultados permite determinar el valor del propio sistema educativo, lo que a su vez señala con claridad qué cambios deberían introducirse para mejorarlo.

No obstante, esta visión mecánica basada en la búsqueda de la eficiencia educativa dista mucho de ser una visión completa y rigurosa del mundo educativo y ha sido muy criticada desde hace bastantes años por diferentes autores, que denominan a dicha perspectiva paradigma técnico o tecnocrático de la educación (Habermas, 1990; Pérez Gómez y Gimeno, 1983; Trillo, 1994; Zeichner y Liston, 1987) y que analizan en

profundidad los sesgos de esta perspectiva educativa[2]. Tal y como afirma Pérez Gómez (1992):

> *Tales explicaciones valen para comprender el funcionamiento de organismos simples, cuya dinámica interna sea tan lineal como pueden ser las primeras reacciones del niño o la niña o, las conductas animales, limitando su análisis a lo observable, es decir, lo que se da fuera de la caja negra. No válidas, sin embargo, para explicar los mecanismos psíquicos más complejos, como los del ser humano, en los que el aprendizaje no se puede entender como una simple relación entre entradas y salidas, ya que éstas son más el producto de la estructura interna que del carácter de la estimulación exterior.* (p. 39)

Que las pruebas estandarizadas operan desde este paradigma técnico basado en la eficiencia y en el conductismo, y que es insuficiente o no explica con claridad lo que acontece en ciencias sociales, es un asunto que ya ha sido tratado por algunos autores que, refiriéndose a PISA, afirman que (Barquín, Gallardo, Fernández, Yus, Sepúlveda y Serván, 2011):

> *No está tan claro que este tipo de pruebas tenga un efecto inmediato sobre la calidad de los sistemas educativos, por cuanto la evaluación está lejos de plantearse como una evaluación educativa o formativa y además su diseño y conclusiones no permiten detectar los fallos del sistema a un nivel de detalle suficiente para poder intervenir. De hecho, en los documentos oficiales de PISA se afirma que "no se pretende medir lo que los alumnos han aprendido". Pero lo cierto es que la proyección internacional que tienen estas pruebas y el efecto revulsivo que provoca, vía medios de comunicación, sobre los Gobiernos de distintos países, son reales y evidentes, a pesar de que desde PISA se reitera que "su intención no es evaluar sistemas educativos". Sencillamente dicen tratar de medir resultados o desempeños, los outs seleccionados por PISA a partir de los inputs realizados por sistemas educativos distintos. Nada sobre lo que ocurre entre ambos extremos, es decir, un modelo de caja negra.* (pp. 325-326)

Por lo tanto cabe plantearse, cuál es la utilidad de estas pruebas que ha propiciado que tengan tanto éxito. Ya que como se viene planteando, las simplificaciones que se desprenden de los resultados de los informes son muy peligrosas ya que los resultados no pueden comprenderse sin los procesos e igualmente como veremos en apartados posteriores, tienen implicaciones claras para la justicia social y la labor de la educación en este sentido.

Atribuimos explicaciones, como exponíamos antes, lineales, simples. Entendiendo que los procesos que dan lugar a tales resultados no son importantes para comprender el cómo y el porqué de lo finalmente observable (en este caso, las respuestas del alumnado en cada una de las pruebas). Se produce una pobre interpretación de la realidad educativa que puede acabar provocando consecuencias importantes en las actitudes y comportamientos de los diferentes agentes del Sistema Educativo, donde la consecución del resultado acaba siendo la meta. De modo que el "ruido" que hacen los resultados de las pruebas estandarizadas, es tan grande que además de crear una determinada opinión pública, nos devuelve a un sistema meramente eficientista, donde lo que importa es la relación entre los "inputs" y los "outpous", olvidándose de analizar la calidad de los

[2] En contraposición al paradigma técnico, se centra en la búsqueda de la calidad de los procesos educativos como camino hacia la mejora educativa. Esta visión, basada en el constructivismo, hace hincapié en la importancia de lo que acontece dentro de la "caja negra" (la mente de los sujetos), como clave fundamental para interpretar el mundo y emitir respuestas, que dejan de mostrarnos la "realidad" tal como promulga la perspectiva conductista.

contextos y los procesos donde tienen lugar los intercambios educativos, atribuyendo unas carencias a los estudiantes de un modo totalmente reduccionista y extrapolando falazmente que la evaluación de competencias del alumnado está suponiendo una medida de detección y mejora de la calidad de sistema educativo.

> *La evaluación que se hace de los alumnos suele explicar el fracaso de estos por su exclusiva responsabilidad [...] No es muy lógico que las atribuciones causales se hagan con tan poco rigor. ¿No hay parte de responsabilidad en el sistema, en el profesor, en el clima, en la forma de evaluar?.* (Santos Guerra, 2014, p. 61)

Es por ello que aludimos a que la aplicación de pruebas estandarizadas parece olvidar que esos resultados no son más que la punta del iceberg y que necesitamos análisis más profundos sobre cuestiones de mayor calado educativo. Sin obviar, por otra parte, aquellos otros asuntos sobre si tales pruebas miden realmente lo que dicen medir (Alcaraz, Caparrós, Soto, Beltrán, Rodríguez y Sánchez, 2013).

3. Los problemas derivados de las pruebas estandarizadas

3.1. La confusión del concepto "evaluación educativa"

Pese a que el término evaluación es, con toda probabilidad, de los más tratados en el ámbito educativo, aún se sigue a día de hoy confundiendo este con otras prácticas y diseños que poco o nada tienen que ver con la idea de evaluación educativa. De hecho, lo que ocurre habitualmente es que (Alcaraz, 2015):

> *En la actualidad aún seguimos en el ámbito educativo confundiendo algunas características de la evaluación con las de la calificación. De modo que, prácticas que dicen ser de evaluación tienen tras de sí muchas de las cualidades propias del acto de calificar. Decimos que evaluamos para comprobar si el alumnado está aprendiendo o no y olvidamos que la principal función de la evaluación no es tanto comprobar el aprendizaje como asegurar las condiciones para que se produzca dicho aprendizaje. Esto, que puede parecer un pequeño matiz, implica tener una serie de consideraciones en el aula que marcarán la diferencia real entre estar poniendo el acento en calificar o estar poniéndolo en evaluar.* (pp. 210-211)

Y es que la evaluación educativa tiene que ver con recoger información útil y veraz, de forma sistemática, para mejorar la práctica educativa (Alcaraz, 2015; Alcaraz, Fernández y Sola, 2012; Álvarez Méndez 1993, 2001, 2007; Casanova, 1992; Elliott, 1990; Fernández Pérez, 2005; Gimeno y Pérez Gómez, 1992; Margalef, 2014; Sola, 1999; Stenhouse, 1987). En el caso de PISA y otras pruebas estandarizadas sí que recogen información de forma sistemática, sin embargo lo que está en cuestión es qué tipo de información se recoge, ya que (Alcaraz et al., 2013):

> *La información proporcionada por estas evaluaciones externas es fundamentalmente acumulativa y, en el mejor de los casos, resulta útil para identificar el progreso de cada escuela o del sistema, pero su utilidad es menor cuando se trata de orientar las prácticas de mejora. Al fin y al cabo, estas pruebas ni diagnostican la naturaleza de los procesos, ni identifican las causas o factores que intervienen en la consecución de ese determinado grado de desarrollo, sus fortalezas y debilidades.* (pp. 581-582)

Esta cuestión, unida a la imposibilidad de medir con exactitud el aprendizaje, si entendemos este desde un punto de vista constructivista, hacen que la evaluación educativa, lejos de enfocar sus esfuerzos en saber qué ha aprendido el alumnado, se deba centrar en recoger información sobre las condiciones en las que se dan la enseñanza y el

aprendizaje. Cuestiones estas que sí son susceptibles de ser valoradas y mejoradas y cuya optimización repercutirá en el aprendizaje del alumnado. Ya que (Alcaraz, 2014):

> *El aprendizaje del ser humano viene determinado por sus esquemas cognitivos y la enseñanza debe ir encaminada a transformar y reconstruir tales esquemas. Frente a la lógica conductista y la pasividad, propia de ésta, en la que el niño o la niña no es más que un recipiente vacío (tabula rasa), aparece, como se viene exponiendo, un nuevo modelo que recoge un principio irrenunciable en la pedagogía, el principio de actividad del sujeto que aprende. Esta concepción dota de una enorme complejidad a las explicaciones sobre el comportamiento humano. Detrás de cada conducta existe un entramado de conexiones que dan lugar a interpretaciones divergentes que dependen exclusivamente del esquema cognitivo que cada ser posee. Añade, por tanto, complejidad al acto de evaluar.* (p. 61)

Es por esto que las pruebas estandarizadas centradas en valorar las competencias adquiridas por el alumnado, se alejan de la intención evaluadora desde el punto de vista educativo que busca la comprensión del proceso de enseñanza-aprendizaje. Principalmente por dos motivos: 1) Las preguntas y problemas que componen dichas pruebas tienen que ver con contenidos, ya sean entendidos como competencias o como contenidos, pero en ningún caso tienen que ver con las condiciones en que se produce el aprendizaje del alumnado, que es lo sustantivo de la evaluación educativa. 2) Sólo el alumnado es objeto de estas pruebas, como si únicamente ellos y ellas tuvieran información útil para mejorar el proceso educativo; ignorando uno de los principales aspectos metodológicos de cualquier proceso evaluativo: la triangulación como metodología de investigación que nos permite contrastar la información y construirla, desde diferentes perspectivas y visiones personales y temporales, de forma que se ajuste más a la realidad.

Con respecto al primer motivo, existen investigaciones que han dejado patente que si bien estas pruebas diagnósticas están bien realizadas y todos los años se va mejorando su diseño, siguen en su mayor parte buscando la competencia reproductora (Alcaraz et al., 2013; Barquín et al., 2011; Gallardo, Fernández, Sepúlveda, Serván y Yus, 2010; Yus, Fernández, Gallardo, Barquín, Sepúlveda y Serván, 2013). Esta no es exactamente una cuestión técnica, no tiene que ver con que haya que mejorar el diseño de dichas pruebas, sino con que los resultados de aprendizaje que las pruebas estandarizadas pueden medir, de ninguna manera equivalen al aprendizaje real que, como decíamos anteriormente, no puede ser medido en tanto en cuanto se trata de modificaciones en la estructura cognitiva del sujeto que aprende. Por lo tanto, por mucha inversión y esfuerzo que dediquemos a perfeccionar una prueba, difícilmente podrá ofrecernos datos exentos de un elevado grado de incertidumbre. Tal y como expone Pérez Gómez (2016):

> *La enorme complejidad que implica la medición del desarrollo humano ha conducido, por lo general, a estrechar el objeto de valoración a lo que es fácilmente medible, a lo que cuesta menos medir.* (p. 10)

Fruto de esta corriente certificadora de aprendizajes tan propia del ámbito educativo, las cuestiones y problemas que se plantean en las pruebas estandarizadas tienen que ver con contenidos, dejando de lado otras cuestiones que afectan al aprendizaje y que podrían influir realmente en su mejora. Es decir, no es fácil encontrar en este tipo de pruebas cuestiones que traten de comprender el tipo de metodología de las clases, los recursos o la calidad de los materiales a disposición del alumnado –evaluación basada en estándares frente a evaluación comprensiva, en el lenguaje empleado por Stake (2006)–; asuntos que sin embargo, sí determinan la calidad de los aprendizajes y sobre los que son necesario deliberar para mejorar la calidad de los sistemas educativos.

Por otra parte, el problema de la evaluación educativa no es de carácter técnico, sino moral y político: los juicios acerca de cómo han de ser mejoradas las condiciones para que se produzcan aprendizajes de mayor trascendencia y calidad tienen que ver con las cualidades de los procesos y de los contextos, no con la perfección de los instrumentos de medición[3].

Con respecto al segundo motivo, este es más evidente aún: las pruebas estandarizadas parecen entender que únicamente el alumnado puede aportar información para la mejora. Cuando la realidad es que de tratarse de una verdadera evaluación, el profesorado y la administración, así como las familias, deberían ser preguntadas acerca de aquellas cuestiones sobre las que puedan arrojar luz y permitirnos por tanto encontrar aspectos que lleven a una mejora real de las condiciones en las que el alumnado aprende.

Esta focalización en el alumnado "extirpa" la posibilidad de realizar una triangulación de la información obtenida, estrategia primordial en todas las investigaciones educativas para contrastar y construir la información obtenida, que se deja absolutamente de lado en estas pruebas estandarizadas en pro de análisis meramente estadísticos cuya validez para comprender fenómenos complejos está más que cuestionada desde hace años.

Por tanto, una vez hemos puesto de manifiesto que las pruebas estandarizadas no se encuentran bajo el paraguas del término evaluación tal y como lo entiende los expertos en la materia, la imposibilidad real de medir tanto competencias como aprendizaje, así como el paradigma técnico desde el que son concebidas, insuficiente a todas luces para explicar los fenómenos complejos que se dan en el área educativa, vamos a exponer aquí una serie de efectos perversos que producen las mismas y que ofrecen, a nuestro parecer, argumentos más que suficientes para dejar de realizarlas en el alumnado.

3.2. El mal uso político de los resultados

Una cuestión que merece la pena recoger aquí aunque no tiene específicamente que ver con la filosofía de estas pruebas, pero que sí se debe contemplar, es el uso político que se les da.

En el caso de España, centrado en las pruebas PISA, los datos ofrecidos se han convertido en armas arrojadizas entre la clase política para justificar las decisiones de las administraciones, a base de reformas educativas que tienen poco de esto último y mucho de ideología enmascarada de neutralidad técnica. Tal y como afirman Saura y Luengo (2015) "en los procesos de legitimación de los discursos políticos se apuesta por determinadas concepciones ideológicas" (p. 143).

Es el caso de la LOMCE[4] (última ley educativa en España) cuya puesta en marcha ha tenido el dudoso honor de ser la primera ley que pone en su contra a colectivos, tradicionalmente con bastantes diferencias entre sí, como son alumnado, familias y profesorado. De nuevo Saura y Luengo (2015) aportan pistas sobre esta estrecha relación:

[3] Aún en el supuesto de que la medida fuese exacta, sin ningún margen de error, la elección entre diferentes alternativas de mejora entra dentro del terreno de la deliberación, no de la traducción mimética de un dato técnico a una actuación predefinida.

[4] Ley Orgánica para la Mejora de la Calidad de la Educación.

> *El breve análisis del discurso político de la LOMCE es solamente un examen efímero para comprender cómo el régimen de la estandarización y las nuevas relaciones de poder y control que ejerce la OCDE en esta gobernanza global están dando lugar a nuevos procesos reformistas de los sistemas escolares.* (p. 143)

Estas pruebas se han convertido en la piedra angular sobre la que se sostiene lo que se ha venido a llamar función política de las reformas educativas (Sola, 1999)[5]. Esta función política de las reformas, donde actualmente tienen mucho peso estas pruebas diagnósticas, es un factor potente que ayuda a comprender por qué las reformas educativas no transforman las prácticas docentes y, por ende, la educación (Cuban, 1990; Fernández Navas, 2015; Fullan, 1993; Gimeno, 1992; Sola, 1999, 2000).

Y es que los datos estadísticos que ofrecen estas pruebas estandarizadas, y las conclusiones superficiales y sesgadas que ofrecen los diferentes grupos de intereses, son fácilmente aceptados por la sociedad sin un análisis riguroso de la validez de los mismos o la idoneidad de los conceptos que tratan de ser representados a través de ellos.

El caso de España, en este sentido es ilustrativo como prueba del uso político que se hace de estas pruebas. Dentro de la propia lógica de las pruebas, se resaltan para crear corrientes de opinión aquellas cuestiones que interesan. Por ejemplo, en España se destacan continuamente en los medios y en los discursos políticos los datos sobre lengua o matemáticas para ilustrar el bajo nivel que ha obtenido el alumnado[6]. No obstante, nunca aparece en ellos que según la propia PISA el sistema educativo español es de los que puntúa más alto en cuanto a equidad del mismo. Cuestión esta que, frente al extendido discurso de la excelencia educativa, debería ser una de las principales preocupaciones de la educación en España. Tal y como se recoge en la Ley Orgánica de Educación, Capítulo 1, artículo 1. Principios:

> *El sistema educativo español, configurado de acuerdo con los valores de la Constitución y asentado en el respeto a los derechos y libertades reconocidos en ella, se inspira en los siguientes principios:*
>
> *a) La calidad de la educación para todo el alumnado, independientemente de sus condiciones y circunstancias.*
>
> *b) La equidad, que garantice la igualdad de oportunidades para el pleno desarrollo de la personalidad a través de la educación, la inclusión educativa, la igualdad de derechos y oportunidades que ayuden a superar cualquier discriminación y la accesibilidad universal a la educación, y que actúe como elemento compensador de las desigualdades personales, culturales, económicas y sociales, con especial atención a las que se deriven de cualquier tipo de discapacidad.* (LOE, 2006, p. 12)

---

[5] Siguiendo la Teoría General de Sistemas enunciada por Bertalanffy (1975), que afirma que el universo físico, biológico y social puede entenderse como un sistema compuesto por subsistemas, cada uno de los cuales es así mismo un sistema, de tal manera que las relaciones entre elementos de uno cualquiera de los subsistemas afecta al sistema en su conjunto, Gimeno (1992) afirma que las reformas educativas hacen hincapié en las relaciones internas (aquellas que se establecen entre los elementos del sistema educativo: profesorado, currículum, ratio, métodos, etc.), en un intento por ocultar los desajustes que se producen en las relaciones externas (aquellas que tienen que ver con la economía, la justicia, el poder… es decir, las relaciones entre subsistemas del macrosistema social, que poco tienen que ver con las condiciones internas del sistema educativo y mucho con la organización social, económica y política de un país).

[6] Cuestión esta que también resulta discutible. Dada las puntuaciones que obtiene el alumnado en PISA, la diferencia, si la extrapolamos a la numeración típica de un examen (de 0 a 10), resulta irrisoria comparada, por ejemplo con Finlandia. Sin embargo, leyendo los medios informativos o los discursos políticos, pareciera que es un desastre absoluto.

Esta primera cuestión desarrollada nos lleva a la segunda: el desmantelamiento de la función social de la escuela.

3.3. El desmantelamiento de la función social de la escuela

Hablar de pruebas estandarizadas nos lleva inevitablemente a hablar del establecimiento de rankings de centros. Tal y como exponen Román y Murillo (2014) bajo el lema de la libertad de elección de centros subyace la idea de que generar competencia entre centros educativos implica un interés por su mejora, haciendo que "sobrevivan aquellos que ofrecen un mejor producto a los consumidores [...] La excusa es que los resultados de tales evaluaciones supuestamente objetivas orientará las decisiones de las familias en el proceso de elección de escuelas" (p. 5). Sin embargo, "sabemos que existe una profunda distancia entre lo que padres aprecian del aprendizaje y las calificaciones de sus hijos y lo que supone y reflejan las evaluaciones estandarizadas del rendimiento escolar" (p. 6).

El hecho es que la proliferación de prácticas de evaluación estandarizada nos aleja del sentido original de la evaluación educativa, ya que los resultados no están sirviendo para la reflexión del profesorado, administradores y políticos (Román y Murillo, 2014). Y nos acerca hacia políticas que ensalzan el discurso de la excelencia y que tanto atentan con la función social de la escuela, puesto que:

> *Las evaluaciones estandarizadas no hacen más que exacerbar una mirada crítica y descalificadora sobre las escuelas y los estudiantes más pobres y vulnerables, quienes también inevitablemente van a ser quienes exhiban los peores resultados según este indicador. Así, el uso de estas mediciones estandarizadas para fomentar la competencia entre centros en un sistema regulado mercantilmente, solo sirve para agudizar la segmentación y hacernos creer que el fracaso escolar es resultado y responsabilidad de los propios estudiantes.* (Román y Murillo, 2014, p. 6)

Por lo tanto, separar los sistemas sociales (económico, político o educativo) como si fueran independientes resulta ser una falacia. La realidad es que todos se interrelacionan y se ajustan y reajustan por influencias externas e internas.

Es en este sentido, en que exponemos que las pruebas estandarizadas penalizan y estigmatizan a la educación pública y benefician claramente a la educación privada (Mistral, 2009). Por definición propia, la educación pública acoge a una amplia diversidad de estudiantes, de ambientes y situaciones socioculturales muy diversos. Esta situación de partida hace que, lógicamente, los resultados en su conjunto en cualquier tipo de prueba sean peores que los de la educación privada, que acoge a estudiantes de ambientes socioeconómicos y culturales altos y muy altos. La razón no estriba necesariamente en una educación de peor calidad en los centros de la red escolar pública, sino en la capacidad que tienen las familias de la red privada para proporcionar a sus hijos e hijas la compensación académica que necesiten, además de la de por sí más elevada cuota de interés por la escuela y de vivencia cultural en general, dada la conocida asociación entre los niveles económico y educativo. Es debido a esta situación de partida que las pruebas estandarizadas ofrecen una visión ilusoria de la "calidad" educativa de ambas instituciones, basada en una comparativa desigual e interesada.

En contra del paradigma que promulga la excelencia educativa y que sólo los y las mejores estudiantes deben progresar en los itinerarios educativos, la realidad es que la educación de más calidad –incluso más excelente podríamos decir también– es aquella que atiende a los y las estudiantes en peor situación de partida y los hace mejores, no la que acoge a los y las mejores estudiantes de partida y los hace progresar algo.

No conviene llevarse a engaños en este aspecto. Es una comparativa desigual, que no se sostiene siquiera bajo parámetros de análisis cuantitativos de datos, y que no tiene que ver con educación. Tiene que ver con justicia social, igualdad de oportunidades, lucha de clases y economía y poder, sobre todo con esto último. Viene al caso recordar los antiguos análisis sobre la función social de la escuela y el acceso al conocimiento que se plantea al alumnado para facilitar -o no- que la escuela siga reproduciendo las clases sociales (Anyon, 1983; Apple, 1991; Baudelot y Establet, 1975; Bowles y Gintis, 1976; Bourdieu y Passeron, 1981; Connell, 1999; Giroux, 2001; Willis, 1988).

3.4. La primacía del "teaching to the test"

En apartados anteriores se exponía una reflexión sobre los peligros que supone el uso interesado de los resultados de las pruebas estandarizadas, así como de la amenaza que supone para la consolidación de un sistema educativo que garantice la igualdad de oportunidades para todo el alumnado. Expondremos ahora otra preocupación que debe suscitar el uso de tales pruebas: Centrar la educación en superar pruebas estandarizadas. Es lo que se ha venido a llamar "teaching to the test" y viene a ser otro de los problemas más importantes derivados del peso que están adquiriendo la estandarización de la evaluación. La educación de calidad es (Pérez Gómez, 2012):

> *Un proceso por el cual estimulamos que cada individuo se interrogue y cuestione el valor y sentido de los esquemas de interpretación, toma de decisiones y actuación, conscientes o tácitos, que ha adquirido a lo largo de su vida. A diferencia de la socialización, la educación supone apostar por el sujeto, por la construcción de la personalidad elegida por cada uno, por potenciar el propio proyecto vital, sin barreras culturales, ideológicas, religiosas o políticas, permitiendo que cada individuo trascienda el escenario concreto y limitado que le ha tocado vivir.* (p. 31)

Aceptando esta definición, vemos que educar es un proceso largo y complejo que no puede darse centrado en la búsqueda de resultados inmediatos para pasar una prueba. Si queremos que el alumnado desarrolle aprendizajes de calidad y competencias de orden superior, es necesario dedicar tiempo al trabajo en nuestras aulas y diseñar espacios y problemas para el alumnado que potencien sus posibilidades de aprendizaje al máximo.

Esto entra directamente en conflicto con la preocupación que despiertan las puntuaciones obtenidas en estas pruebas estandarizadas para el profesorado que, preocupado por el desprestigio que supone obtener una mala puntuación para su centro, empieza a centrarse en "entrenar" a su alumnado para obtener mejores resultados[7].

[7] Una vez más, no hay nada nuevo en esta afirmación. En España es de sobra conocido que el segundo curso de Bachillerato (el último de la Educación Secundaria) se ha convertido mayoritariamente en un curso preparatorio de la prueba de acceso a la universidad (PAU) debido a la preocupación del profesorado por que su alumnado apruebe y consiga buena nota. La manera menos problemática de intentar ese resultado deseado es anular por completo todo un curso académico (desperdiciar la posibilidad de realizar nuevos y valiosos aprendizajes) y tratar de entrenar a los estudiantes en los contenidos y en las preguntas frecuentes de la PAU, mimetizando la actividad académica con la que ordena el ritual del examen de entrada a la institución universitaria.

Con la aparición de las pruebas estandarizadas estamos a un paso de convertir todo el sistema educativo en un sistema de preparación de exámenes para salir airosos de ellos, aunque en el proceso se pierda la virtualidad educativa y se renuncie a una enseñanza y aprendizaje de calidad.

Lo cual deja de lado las posibilidades no sólo de acercar a su alumnado a contenidos más relevantes, sino de hacerlo a través de procesos ricos, que podrían llevar a estos a construir aprendizajes relevantes y desarrollar competencias de orden superior.

Igualmente, aleja al profesorado de emprender procesos ricos de innovación educativa que, preocupado por la mejora de las puntuaciones de sus estudiantes, olvida que la importancia de la educación y los aprendizajes se encuentra en la calidad de los procesos y no en las puntuaciones.

Esto supone, al igual que otros viejos problemas para el profesorado como la burocratización de su trabajo, un peligro real de alienación del pensamiento docente (Contreras, 2011) que, centrado en la obtención de resultados inmediatos –los de las pruebas, verdadera espada de Damocles– no tiene tiempo para pensar cómo desarrollar procesos ricos en actividades de calidad que permitan al alumnado obtener aprendizajes valiosos.

Por otro lado, merece la pena rescatar aquí una cuestión que se hace más que evidente con las pruebas estandarizadas. Se trata de la preocupación de estar continuamente "examinando" al alumnado para ver qué ha aprendido o cómo va aprendiendo, en lugar de preocuparnos por cómo diseñar espacios y problemas de calidad para proporcionar más y mejores oportunidades de aprendizaje. Preocupación esta bajo la que operan, y que cala subrepticiamente en el pensamiento del profesorado y de la sociedad en general en el profesorado y la sociedad, las pruebas estandarizadas.

3.5. El privilegio de las materias instrumentales

Otro de los problemas que presentan el uso de estas pruebas estandarizadas es que se centran únicamente en determinadas disciplinas. En palabras de Hernández (2015, p. 126): "Lo que PISA evalúa es lo que debe ser enseñado". El autor expone que uno de los efectos colaterales de estas pruebas es el de considerar como "conocimiento válido para aprender aquel que es más fácil de colocar dentro de la perspectiva de transferibilidad que plantean dichas pruebas (Hernández, 2015).

A esta problemática hace referencia Robinson (2006), cuando afirma que por distintos que sean los sistemas educativos, todos los del mundo entero están estructurados de tal forma que dan más importancia a las materias científicas, luego a las humanidades y después a las artes. Y que ello es debido a que la escuela se pensó para enseñar los conocimientos necesarios para el mundo del trabajo que surge inmediatamente después de la revolución industrial.

Y es que esta tradición escolar, de la que no escapan las pruebas estandarizadas, no hace más que poner de manifiesto para alumnado, profesorado y sociedad que existe una jerarquía de disciplinas en función de su importancia. Y que son, tradicionalmente, las ciencias las más beneficiadas y las artes las más perjudicadas.

> *Cabe preguntarse a qué responde que se evalúen sólo Matemáticas, Ciencias, Lengua y ahora la competencia digital y las habilidades diarias. Materias que por cierto, son las que contribuyen y responden a las expectativas economicistas que impulsa y favorece esta organización internacional*[8]. (Hernández, 2015, p. 126)

[8] Se refiere a la Organización Internacional para la Cooperación y el Desarrollo Económico (OCDE).

Es importante recordar que si bien las necesidades de la industrialización requerían la formación acelerada de mano de obra en los rudimentos del cálculo, la lectura y la escritura, parece evidente que las condiciones de la Modernidad Líquida (Baumann, 2007) deberían exigir otros dominios, otras destrezas, otros conocimientos que favorezcan la formación inter y multidisciplinar que requieren todos los sistemas educativos entre cuyas pretensiones se encuentra la de formar ciudadanos y ciudadanas que sepan desenvolverse en la sociedad que los rodea. Y en esta cuestión (una de las finalidades en última instancia de la educación) juegan un papel fundamental todas las materias; no sólo las instrumentales.

A este respecto se pronuncia igualmente Apple (2002) cuando afirma que:

> *Los neoconservadores lamentan el «abandono» del currículo tradicional y de la historia, la literatura y los valores que dicen representaba. Tras esta inquietud se encuentran unos supuestos históricos sobre «la tradición», la existencia de un consenso social en torno a lo que se debe considerar conocimiento legítimo y la superioridad cultura*. (p. 67)

4. Conclusiones

Como se puede observar a lo largo de este trabajo, las pruebas estandarizadas si algo tienen asociado, es la controversia que suscitan.

Es por ello que nos parecía necesario hacer una reflexión desde una perspectiva crítica de aquellos aspectos que, a nuestro juicio, deben ser analizados y valorados desde investigaciones de forma rigurosa.

Para nosotros parece importante destacar, en primer lugar la imposibilidad y también porqué no la idoneidad, de una medición exacta –si no es exacta no es medición, más bien hablaríamos de intuiciones poco precisas sobre los aprendizajes del alumnado. Tal y como se expuso en los inicios de este trabajo.

Por otro lado, parece evidente que estas pruebas tienen poco o nada que ver con la evaluación educativa; en primer lugar porque sólo prestan atención a los aprendizajes del alumnado. Dejando de lado otras cuestiones que también tienen que ver con la calidad de los sistemas educativos y que inciden en el aprendizaje del alumnado –algunas de ellas, las que más influyen en ellos, como la calidad del profesorado tal que afirman Hattie (2008, 2011) y Delors (1996)–.

En segundo lugar, porque este tipo de pruebas, al centrarse únicamente en este aspecto, dejan de lado la posibilidad de la triangulación de datos. Aspecto esencial para que una evaluación –y cualquier investigación– sea rigurosa.

Que este tipo de pruebas deje de lado aspectos tan importantes como la triangulación tiene que ver con el paradigma desde el que se sitúan. Tal que tratábamos de exponer en este trabajo, la idea bajo la que opera la lógica de dichas pruebas, tiene que ver con un paradigma conductista bajo el cual el estudio estadístico puede reflejar cuestiones complejas como el aprendizaje masivo del alumnado y, por ende, la calidad del sistema educativo en el que lo hacen.

Nada más lejos de la realidad, existe un amplia crítica en la literatura pedagógica y de investigación cualitativa sobre las limitaciones de este tipo de estudios para explicar fenómenos complejos como puede ser el acto educativo. No obstante, parece que este asunto se olvida a la hora de poner en práctica este tipo de pruebas. Quizás, como

tratábamos de explicar con anterioridad, por lo valioso de los datos que ofrecen, para calar en la opinión pública y, por tanto, convertirse en armas arrojadizas entre las diferentes administraciones.

En último lugar, tratábamos de exponer, los graves problemas que se derivan de la proliferación de este tipo de pruebas:

Por un lado, está el asunto de la Justicia Social y la igualdad de oportunidades, y es que si existe un sector beneficiado, de forma generalizada, de los resultados de estas pruebas es la enseñanza privada que independientemente de su calidad, sale más beneficiada que la educación pública. Ya que esta última atiende a todo el alumnado, independientemente de su procedencia. Mientras que la privada, por el tipo de perfil de las familias a las que atiende, tiene en cierta forma, asegurada la preocupación académica de las mismas.

Por otro lado, está lo que hemos llamado: El problema del foco de atención. Y es que estas pruebas, por la importancia y repercusión que están teniendo sus resultados, están desplazando la atención de alumnado, familias y docentes –sobre todo estos últimos– de la preocupación por realizar actividades educativas y de calidad con su alumnado hacia el entrenamiento para obtener mejores puntuaciones en dichas pruebas.

Por último, encontramos el asunto de las disciplinas. Por desgracia en la tradición escolar siempre se ha asentado la idea de que ciertas áreas de conocimiento son más importantes que otras. Las pruebas estandarizadas no son ajenas a esta cultura institucional y se centran en aquellas disciplinas que son consideradas por las mismas, más importantes. Aumentando, si cabe, la escasa dedicación y preocupación por disciplinas como el arte, la música, etc. De las que numerosos autores y autoras han destacado su importancia central para la educación en mayúsculas.

Referencias

Alcaraz, N. (2014). Un viejo trío de conceptos: Aprendizaje, currículum y evaluación. *Aula de Encuentro*, *2*(16) 55-86.

Alcaraz, N. (2015). Evaluación versus calificación. *Aula de Encuentro*, *2*(17), 209-236.

Alcaraz, N., Caparrós, R., Soto, E., Beltrán, R., Rodríguez, A. y Sánchez, S. (2013). ¿Evalúa PISA la competencia lectora? *Revista de Educación*, *360*, 577-599. doi:10.4438/1988-592X-RE-2011-360-130

Alcaraz, N., Fernández, M. y Sola, M. (2012). La voz del alumnado en los procesos de evaluación docente universitaria. *Revista Iberoamericana de Evaluación Educativa*, *2*(5). Recuperado de http://www.rinace.net/riee/numeros/vol5-num2/art2_htm.html

Álvarez Méndez, J. M. (1993). El alumnado. *Cuadernos de Pedagogía*, *219*, 28-32.

Álvarez Méndez, J. M. (2001). *Evaluar para conocer, examinar para excluir*. Madrid: Morata.

Álvarez Méndez, J. M. (2007). La evaluación formativa. *Cuadernos de Pedagogía*, *364*, 96-101.

Apple, M. (1991). *Ideología y currículo*. Madrid: Akal.

Apple, M. (2002). *Educar como Dios manda*. Barcelona: Paidós.

Angulo, J. F. (1999). Entrenamiento y 'coaching': Los peligros de una vía revitalizada. En D. F. Labaree, J. Smyth, K. M. Zeichner, St. Kemmis, A. Hargreaves, J. Elliott y M. Cochran-Smith (Eds.), *Desarrollo profesional del docente: Política, investigación y práctica* (pp. 467-505). Madrid: Akal.

Anyon, J. (1983). Intersections of gender and class: Accommodation and resistance by working-class and affluent females to contradictory sex- role ideologies. En S. Walker y L. Barton (Eds.), *Gender, class, and education* (pp. 19-38). Londres: Falmer.

Barquín, J., Gallardo, M., Fernández, M., Yus, R., Sepúlveda, M. y Serván, M. (2011). Todos queremos ser Finlandia. Los efectos secundarios de PISA. *Education in the Knowledge Society (EKS), 12*(1), 320-339.

Baudelot, C. y Establet, R. (1975). *La escuela capitalista.* Madrid: Siglo XXI.

Bauman, Z. (2007). *Los retos de la educación en la modernidad líquida.* Barcelona: Gedisa.

Bertalanffy, L. (1975). *Perspectives on General Systems Theory. Scientific-philosophical studies.* Nueva York, NY: George Braziller.

Bourdieu, P. y Passeron, J. C. (1981). *La reproducción. Elementos para una teoría del sistema de enseñanza.* Barcelona: Laia.

Bowles, S. y Gintis, H. (1976). *La meritocracia y el coeficiente de inteligencia: Una nueva falacia del capitalismo.* Barcelona: Anagrama.

Casanova, M. A. (1992). *La evaluación, garantía de calidad del centro educativo.* Zaragoza: Edelvives.

Connell, R. (1999). *Escuelas y justicia social.* Madrid: Morata.

Contreras, J. (2011). *La autonomía del profesorado.* Madrid: Morata.

Cuban, L. (1990). Reforming again, again and again. *Educational Researcher, 19*(1), 3-13. doi:10.2307/1176529

Delors, J. (1996). *La educación encierra un tesoro.* París: Unesco.

Elliott, J. (1990). *La investigación-acción en educación.* Madrid: Morata.

Fernández Navas, M. (2016). Universidades, enseñanza virtual, competencias y justicia social: Una historia de mercantilización de la enseñanza superior. *Aula de Encuentro, 1*(18), 251-273.

Fernández Navas, M. (2015). *Internet, organización en red y educ@ción. Estudio de un caso de buenas prácticas en enseñanza superior* (Tesis doctoral). Universidad de Málaga, Málaga.

Fernández Pérez, M. (2005). *Evaluación y cambio educativo.* Madrid: Morata.

Fullan, M. (1993). *Change forces. Probing the depths of educational reform.* Nueva York, NY: The Falmer Press.

Gallardo, M., Fernández, M., Sepúlveda, P., Serván, M. J. y Yus, R. (2010). PISA y la competencia científica: Un análisis de las pruebas de PISA en el área científica. *RELIEVE, 16*(2). Recuperado de: http://www.uv.es/RELIEVE/v16n2/RELIEVEv16n2_6.htm

Gimeno, J. (1992). Reformas educativas. Utopía, retórica y práctica. *Cuadernos de Pedagogía, 209*, 62-68.

Gimeno, J. y Pérez Gómez, Á. I. (1992). *Comprender y transformar la enseñanza.* Madrid: Morata.

Giroux, H. (2001). *Theory and resistance in education: Towards a pedagogy for the opposition.* South Hadley, MA: Bergin & Garvey.

Habermas, J. (1990). *Conocimiento e interés.* Madrid: Taurus.

Hargreaves, A. y Dawe, R. (1990). Paths of professional development: Contrived collegiality, collaborative culture, and the case of peer coaching. *Teaching and Teacher Education, 3*(6), 227-241. doi:10.1016/0742-051x(90)90015-w

Hattie, J. A. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Nueva York, NY: Routledge.

Hattie, J. A. (2011). *Visible learning for teachers: Maximizing impact on learning*. Nueva York, NY: Routledge.

Hernández, F. (2015). Las materias que distraen o la utilidad de lo inútil. En J. Gimeno (Comp.), *Los contenidos. Una reflexión necesaria* (pp. 125-137). Madrid: Morata.

Margalef, L. (2014). Evaluación formativa de los aprendizajes en el contexto universitario: Resistencias y paradojas del profesorado. *Educación XX1, 17*(2), 35-55. doi:10.5944/educxx1.17.2.11478

Martínez Rizo, F. (2001). Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate. *Revista de la Educación Superior*, *120*(30).

Martínez Rodríguez, J. B. y Sola, M. (2004). *Diagnóstico del sistema nacional de evaluación de la educación en El Salvador*. Granada: Arial.

Ley Orgánica 2/2006, de 3 de mayo, de Educación. Boletín Oficial del Estado, núm. 106, de 4 de mayo de 2006, pp. 1 a 107.

Mistral, G. (2009). Estandarización educativa en Chile: Un peligroso hábito. *Revista Docencia*, *38*, 4-17.

Pérez Gómez, A. (2012). *Educarse en la era digital*. Morata: Madrid.

Pérez Gómez, A. (Dir.). (2016). *El portafolios educativo en Educación Superior*. Madrid: Akal.

Pérez Gómez, A. y Gimeno, J. (1983). *La enseñanza: Su teoría y su práctica*. Madrid: Akal.

Román. M. y Murillo, F. J. (2014). Uso de los resultados de las evaluaciones estandarizadas como criterio de elección y selección de escuelas. *Revista Iberoamericana de Evaluación Educativa*, *7*(1), 5-7.

Santos Guerra, M. A. (2014). *La evaluación como aprendizaje. Cuando la flecha impacta en la diana*. Madrid: Narcea.

Saura, G. y Luengo, J. (2015). Política global más allá de lo nacional. Reforma educativa (LOMCE) y el régimen de estandarización (OCDE). *Bordón*, *67*(I), 135-148. doi:10.13042/Bordon.2015.67109

Sola, M. (1999). El análisis de las creencias del profesorado como requisito de desarrollo profesional. En A. Pérez, J. Barquín y F. Angulo (Eds.), *Desarrollo profesional del docente. Política, investigación y práctica* (pp. 661-683). Madrid: Akal.

Sola, M. (2000). Los efectos de la reforma en la condición profesional de los docentes: La formación de creencias ideológicas y su influencia en el pensamiento profesional. En I. Rivas (Coord.), *Profesorado y reforma: ¿Un cambio en las prácticas de los docentes?* (pp. 73-80). Málaga: Aljibe.

Stake, R. E. (2006). *Evaluación comprensiva y evaluación basada en estándares*. Barcelona: Graó.

Stenhouse, L. (1987). *La investigación como base de la enseñanza*. Madrid: Morata.

TED. (2009, 3 de agosto). *Sir Ken Robinson: Las escuelas matan la creatividad TED 2006* [Archivo de vídeo]. Recuperado de https://www.youtube.com/watch?v=nPB-41q97zg

Trillo, F. (1994). El profesorado y el desarrollo curricular: Tres estilos de hacer escuela. *Cuadernos de pedagogía*, *228*, 70-74.

Villar Angulo, L. M. (1977). *La formación del profesorado: Nuevas contribuciones*. Madrid: Aula XXI.

Willis, P. (1988). *Aprendiendo a trabajar: Cómo los chicos de clase obrera consiguen trabajos de clase obrera*. Madrid: Akal.

Yus, R., Fernández, M., Gallardo, M., Barquín, J., Sepúlveda, P. y Serván, M. J. (2013). La competencia científica y su evaluación. Análisis de las pruebas estandarizadas de PISA. *Revista de Educación*, *360*(enero-abril), 557-576.

Zeichner, K. (1993). El maestro como profesional reflexivo. *Cuadernos de Pedagogía*, *220*, 44-49.

Zeichner, K. M. y Liston, D. (1987). Teaching students teachers to be reflect. *Harvard Educational Review*, *57*(1), 23-49. doi:10.17763/haer.57.1.j18v7162275t1w3w

Breve CV dos autores

Manuel Fernández Navas

Es profesor en el Departamento de Didáctica y Organización Escolar de la Universidad de Málaga y forma parte del grupo de investigación HUM-311 de la Universidad de Málaga. Es maestro de Audición y Lenguaje, licenciado en Psicopedagogía y doctor por la Universidad de Málaga. Sus líneas de investigación giran en torno a la enseñanza virtual, la investigación-acción, las nuevas tecnologías en educación, la formación de docentes y la evaluación educativa. Ha participado en proyectos de evaluación externa de la Consejería de Educación (*EVACENPRO: La evaluación externa de los centros de formación del profesorado de Andalucía)* y del Centro de Estudios Andaluces (*El plan de apertura de centros de Andalucía, en el marco de las nuevas tendencias sociales. Situación y perspectiva*). Es autor de artículos sobre evaluación y formación y coordinador del libro: *La innovación educativa: Más allá de la ficción*, de la editorial Pirámide. ORCID ID: 0000-0002-9445-2643. Email: mfernandez1@uma.es

Noelia Alcaraz Salarirche

Es profesora en el Departamento de Didáctica y Organización Escolar de la Universidad de Málaga y forma parte del grupo de investigación HUM-311 de la Universidad de Málaga. Es licenciada en Pedagogía y doctora por la Universidad de Málaga. Sus líneas de investigación giran en torno a la innovación educativa, la investigación-acción, la formación del profesorado y la evaluación educativa. Ha participado en diferentes proyectos I+D, así como en los proyectos "Evaluación de los Centros de Profesorado Andaluces" y "Evaluación del Plan de Apertura de Centros en Andalucía". Es autora de artículos relacionados con la formación y la evaluación y coordinadora del libro: *La innovación educativa: Más allá de la ficción*, de la editorial Pirámide. ORCID ID: 0000-0002-5296-5248. Email: noe@uma.es.

Miguel Sola Fernández

Es profesor titular en el Departamento de Didáctica y Organización Escolar de la Universidad de Málaga. Doctor por la Universidad de Málaga. Forma parte del grupo de investigación HUM-311. Sus intereses actuales giran en torno a la innovación de la docencia universitaria, las redes sociales como instrumentos para la investigación

naturalista y la evaluación educativa. Es miembro del comité de redacción de *Educational Action Resarch Journal*, ha coordinado las evaluaciones externas de la Educación Secundaria Obligatoria y los Centros TIC andaluces. Es autor de publicaciones relacionadas con la formación, la innovación, la reflexión y la evaluación. ORCID ID: 0000-0001-9195-1597. Email: misola@uma.es

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 69-87.
https://doi.org/10.15366/riee2017.10.1.004

Los Efectos Adversos de una Evaluación Nacional sobre las Prácticas de Enseñanza de las Matemáticas: El Caso de SIMCE en Chile

The Effects Adverse of a National Assessment on the Teaching Practices in Mathematics: The Case of SIMCE in Chile

Carolina Ruminot Vergara *
Universidad de Ottawa

Este artículo es un estudio sobre los efectos de un sistema estandarizado de evaluación la enseñanza de matemáticas, que considera el caso especial de la evaluación SIMCE en Chile. Se ha demostrado que las evaluaciones estandarizadas influyen de manera creciente en los sistema educativos, notablemente sus organizaciones, los programas de estudio, las tareas que se ofrecen, y como consecuencia las prácticas de enseñanza (Clarke, 2013; Cox, 2003; Estudio EMF, 2009). Diversos estudios muestran que los efectos de evaluaciones estandarizadas sobre sistemas educativos no son necesariamente positivos (Mons, 2009; Schoenfeld, 2007). Ellos también muestran que la presión ejercida sobre las escuelas y los profesores para mejorar sus resultados tienden a inducir una concentración perjudicable de la enseñanza. El presente artículo presenta las diversas medidas puestas en marcha para preparar a los estudiantes para la evaluación SIMCE, y pone en evidencia algunos efectos sobre las prácticas de enseñanza, tales como la contracción del currículo a nivel de los contenidos y de las tareas. El presente artículo se enmarca en un trabajo de tesis doctoral que incluye un trabajo de campo en 12 instituciones educativas en Santiago de Chile con 13 profesores pertenecientes a estas instituciones.

Palabras Clave: Evaluación estandarizada, SIMCE, Enseñanza de matemáticas, Efectos de SIMCE.

This article examines the effects of standardized assessment system on mathematics teaching, with a focus on the case of the SIMCE evaluation in Chile. It has been shown that standardized testing has an increasing influence on educational systems, in particular on their organization, their syllabi, the math problems given, and by consequence the teaching practise (Clarke, 2013; Cox, 2003; EMF Study, 2009). Various studies show that the effects of standardized testing on educational systems are not necessarily positive (Mons, 2009; Schoenfeld, 2007). They also show that the pressure placed upon schools and teachers to improve results tend to induce a notable concentration of the teaching in order to prepare students for those tests. This article shows that diverse measures are implemented in order to prepare students for the SIMCE evaluation, while demonstrating some of the effects on the teaching practise, such as the curricular contraction of the contents and the associated problems. This article is based upon a doctoral theses study, which includes a field study of 12 educational institutions in Santiago, Chile, with 13 teachers from those institutions.

Keywords: Standardized Assessment, SIMCE, Mathematics instruction, Effects of SIMCE.

*Contacto: caruminot@gmail.com

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 7 de julio de 2016
1ª Evaluación: 13 de octubre de 2016
Aceptado: 12 de diciembre de 2016

1. Introducción y problemática

Durante las dos últimas décadas hemos asistido a la multiplicación de las evaluaciones estandarizadas a gran escala, tanto a nivel nacional como internacional. Estas evaluaciones influyen cada vez más en los sistemas educativos, e influencian, por un lado, los programas de estudio, sus contenidos, su organización, las tareas propuestas en las evaluaciones y, por otro lado, las prácticas de enseñanza. Diversos estudios muestran que estos efectos no son necesariamente positivos (véase, por ejemplo, Mons, 2009; Schoenfeld, 2007;). Ellos refieren a cómo la presión ejercida sobre los establecimientos escolares y los profesores –para que ellos mejoren sus resultados– produce un resultado adverso en el proceso de enseñanza; resultado que se traduce en una preparación excesiva de los estudiantes a estas evaluaciones ("*teaching to the test*") afectando especialmente a las poblaciones estudiantiles más vulnerables.

En este contexto, viniendo de un país –Chile– donde existe desde 1988 una evaluación estandarizada –SIMCE– realizada por todos los estudiantes de 4º, 6º y 8º Año de enseñanza básica, y de 2º y 3r año de la enseñanza secundaria (con un aumento constante no solo de los niveles de enseñanza en los que se aplica, sino también de las disciplinas que se evalúan) y que desempeña un papel cada vez más importante en la dirección del sistema educativo, se desea a través de este artículo mostrar el impacto de esta evaluación en la enseñanza de las matemáticas en Chile.

La evaluación SIMCE, cuyas siglas significan Sistema de Medición de la Calidad de la Educación, se define como un sistema de evaluación de la calidad de la enseñanza. Como lo señala el informe de la OCDE de 2004 sobre la Revisión de las Polítcas Nacionales de Educación Chilena (OCDE, 2004):

> *La prueba SIMCE en 1988, muchos la vieron como una medida de la "efectividad" de las escuelas y continúan viéndola así. El nuevo gobierno, a su vez, usó los resultados del SIMCE a comienzos de los años 90 principalmente para identificar establecimientos de bajo rendimiento para invertir recursos adicionales en ellas y monitorear el efecto de esas inversiones. Más recientemente, el SIMCE ha sido usado como una medida principal de la calidad y mejoramiento de las escuelas en Chile.* (p. 163)

La evaluación SIMCE ha evolucionado desde su creación y se presenta actualmente de la siguiente manera en el sitio del Ministerio de Educación (2012):

> *Desde 2012, el SIMCE se convirtió en el sistema de clasificación utilizado por la Agencia de Educación de Calidad para evaluar las instituciones de aprendizaje resultados, medir el logro de los contenidos y habilidades del plan de estudios actual en diferentes materias o áreas de aprendizaje, a través de una medida que se aplica a todos los estudiantes del país matriculados en los grados evaluados.* (p. 1)

Estudiar el impacto de SIMCE no es una tarea fácil, porque los efectos de un tal sistema son a priori múltiples, a la vez directos e indirectos, como lo expresa Bodin (2006) cuando señala que:

> *El impacto directo de estos estudios se centran escencialmente en modificaciones de los programas de estudio, la formación docente y las instrucciones qui puedan dar a los profesores.* (p. 80)

Sin embargo, existen efectos menos evidentes a identificar, como lo son las prácticas educativas y los aprendizajes de los estudiantes. Estos efectos, además no son homogéneos a través de un país y de sus diferentes establecimientos. Esos efectos se inscriben asimismo en una dinámica de múltiples determinantes lo que no es fácil de

entender. Mediante este artículo la principal pregunta que se quiere responder es la siguiente: Conociendo la importancia que tiene para las instituciones educativas los resultados de los alumnos a esta evaluación ¿Cuáles son los dispositivos eventualmente puestos en marcha para preparar a los estudiantes y mejorar los resultados? ¿Se observa, en particular, una concentración de la enseñanza entorno a los contenidos evaluados y a los tipos de tareas propuestos en la evaluación?

Por razones operacionales esta investigación se limita a un nivel escolar, el de octavo año de enseñanza (estudiantes de 13-14 años), y al dominio matemático de la geometría, particularmente las magnitudes y la medición. Estas decisiones tienen razones bien precisas. El nivel de la enseñanza escogido, corresponde al último anño de la educación básica, uno de los niveles que SIMCE evalúa. Este nivel posee una historia substancial, y su elección favorece la comparación con evaluaciones internacionales como PISA y TIMSS. Con respecto a la geometría, se trata de un dominio particularmente problemático para los profesores de la educación obligatoria en Chile (Castela, Consigliere, Guzman, Houdment, Kuzniaz y Rauschera, 2006; Ruminot Vergara, 2014).

A través de este artículo se presentan algunos resultados del trabajo doctoral de la autora, el cual tiene como intensión mostrar los efectos explícitos e implícitos del impacto de la evaluación nacional SIMCE sobre las prácticas de enseñanza. Se pudo constatar a través del estudio en terreno, que involucró establecimientos escolares, sus profesores y observaciones de clases, ciertas acciones institucionales claramente orientadas a mejorar los resultados de los estudiantes en la evaluación SIMCE. Asimismo en la puesta en marcha de las prácticas docentes también se observan acciones pedagógicas influenciadas por dicha evaluación.

2. Marco Teórico

El marco teórico que se ha utilizado principalmente es la *"Teoría Antropológica de lo Didáctico"* –TAD– desarrollado por el investigador francés Yves Chevallard (1999). Esta elección tiene una razón principal. Dado que el objetivo del estudio es la comprensión de los efectos de la evaluación nacional SIMCE sobre la enseñanza de las matemáticas en Chile. Para comprender estos efectos se necesita una perspectiva amplia sobre los procesos de ensenñanza y aprendizaje que permitan identificar los diferentes determinantes que pesan sobre un tal sistema de evaluación. Por su enfoque institucional, por la atención que presta a los sistemas de coerciones y condiciones que a diferentes niveles afectan los procesos de enseñanza y aprendizaje a través la jerarquía de niveles de co-determinación didáctica (Chevallard, 2002).

2.1. Noción de praxeología

La TAD formula que toda actividad humana consiste en (Chevallard, 2002):

> *Para llevar a cabo una tarea t de algún tipo T, utilizando algún técnica τ, justificada por una tecnología θ que, que al mismo tiempo es pensanda o incluso producida y, que a su vez se justifica por la teoría Θ. En resumen, toda la actividad humana implementa una organización se pueden señalar como [T, τ, θ, Θ] y nombrar por praxeología, u organización praxeológica.* (p. 1)

Esta noción de praxeología es un elemento fundamental de la TAD. Se utilizó esta noción, tanto para estudiar las actividades matemáticas como las actividades didácticas en las diversas instituciones. Tomando en cuenta el hecho que la actividad matemática se

inscribe en estructuras de niveles de complejidad diferentes, las organizaciones matemáticas expresan este fenómeno mediante diferentes niveles de organizaciones matemáticas, puntuales, locales, regionales y globales. Las organizaciones matemáticas puntuales (OMP), son aquellas en donde se encuentra un único tipo de tarea T. Esta noción es relativa a la institución considerada y está definida, en principio, a partir del bloque práctico-técnico [T/τ]. Las organizaciones matemáticas locales (OML) son el resultado de la integración de un conjunto de OMP.

2.2. Momentos del estudio

En la TAD, el modelo praxeológico se aplica a todas las prácticas humanas y en particular a las prácticas didácticas. Se movilizó este modelo praxeológico a las dimensiones de la investigación que dan acceso directamente o indirectamente a prácticas didácticas: el análisis del programa de estudio y las observaciones de clases. Por ejemplo el programa oficial de estudio deja en evidencia la intención de orientar las organizaciones didácticas, dado que además de presentar los contenidos organizados por temas, explicita coómo el profesor debería desarrollarlos.

La TAD postula que el proceso de estudio visto como construcción o reconstrucción de praxeologías matemáticas posee invariantes que se expresan en el modelo de los momentos de estudio (Chevallard, 2002). Se postula que cada proceso de estudio de una praxeología matemática puntual [T, τ, θ, Θ] incluye los seis momentos siguientes:

- ✓ *El momento de primer encuentro con el tipo de tarea T.* Es el encuentro o re-encuentro con una organización matemática, a través de tipos de tareas constitutivos de la organización matemática en cuestión.
- ✓ *El momento de exploración del tipo de tareas T y de emergencia de la técnica τ.* Es el primer momento como una técnica embrión que evoluciona a la técnica relativamente rutinaria que es justificada por un discurso tecnológico-teórico y que permite resolver problemas asociados a un mismo tipo de tareas.
- ✓ *El momento de construcción del bloque tecnológico-teórico.* Desde el primer encuentro con un tipo de tarea, se establece generalmente una relación con el entorno tecnológico-teórico anteriormente elaborado, o con gérmenes de un entorno por crear que se precisará mediante una relación dialéctica con la emergencia de la técnica.
- ✓ *El momento de trabajo de la praxeología matemática, y en particular de la técnica.* Este momento es donde la técnica es mejorada y, además su puesta en práctica evidencia su eficacia y fiabilidad.
- ✓ *El momento de institucionalización.* Este momento tiene como objetivo precisar exactamente la organización matemática elaborada; validando y excluyendo algunos elementos ya trabajados.
- ✓ *El momento de evaluación.* En esta etapa se observa y se reflexiona sobre la utilidad de lo aprendido y sobre la importancia de conservar este conocimiento para aplicaciones futuras (Chevallard, 1999).

Como lo aclara Chevallard (1999), este orden no corresponde necesariamente al orden en que se encuentran los momentos en una organización didáctica dada. Cada momento no corresponde necesariamente a una unidad temporal única. Ciertos momentos pueden ser

distribuidos sobre varias sesiones. Las sesiones de clase observadas deberían hacer visibles diferentes momentos de estudio. Se interroga sobre las características de las interacciones didácticas en cada tipo de sesión, y en particular si estas características permiten identificar en las prácticas didácticas efectos de la evaluación SIMCE.

3. Metodología

Dado que este artículo se inscribe un trabajo de tesis es oportuno en este punto entregar un panorama general de la investigación realizada, de este modo, situar y delimitar el trabajo que a continuación se presenta. La investigación global incluyó varias fases de estudio:

- ✓ Un estudio comparativo entre las diferentes evaluaciones a gran escala en las que el sistema educativo chileno participa en matemáticas: PISA, TIMSS y LLECE
- ✓ Un análisis de programas del estudio1 de la geometría de 8º grado de enseñanza primaria,
- ✓ Un análisis de los documentos oficiales y dos manuales escolares (un manual oficial y otro manual ampliamente utilizado),
- ✓ Un estudio de doce establecimientos escolares de características socioeconómicas diversas y que declaran explícitamente realizar acciones para mejorar sus resultados en la evaluación SIMCE,
- ✓ La elaboración y aplicación de cuestionarios destinados a profesores de matemáticas,
- ✓ La preparación y realización de entrevistas semi-dirigidas con los profesores de matemáticas y con diferentes actores de esos establecimientos,
- ✓ Observaciones de sesiones de enseñanza, a la vez de ordinarias y de preparación a la evaluación SIMCE en nueve de las instituciones estudiadas.

Por razones de pertinencia en este presente artículo se limitarán los detalles de los diferentes estudios mencionados en el párrafo anterior y se presentarán los principales resultados obtenidos durante el trabajo de campo. Esto incluye el estudio de las instituciones educativas, las entrevistas semi-guiadas con los profesores de matemáticas de estas instituciones y las observaciones de clases. Doce instituciones educativas fueron estudiadas. Mediante entrevistas a los directivos y recolección de información emitida por el Ministerio de Educación, se determinaron las acciones en relación a la evaluación SIMCE. Las entrevistas con los profesores abarcaron cuatros dimensiones de análisis: Formación y experiencia, Posicionamiento institucional, Relación a la evaluación SICME, Sensibilidad Didáctica. Las observaciones de clases fueron realizadas tanto en sesiones ordinarias como de preparación SIMCE. Se observaron en total 17 clases y se

[1] Este programa (Decreto 220-2002) es una herramienta diseñada por el Ministerio de Educación con el objetivo de organizar y orientar el trabajo escolar en función de Objetivos Fundamentales (OF), de Contenidos Mínimos Obligatorios (CMO) y de Aprendizajes Esperados (AE) que se encuentran definidos en el currículo oficial.

consideraron para el estudio 9, correspondientes a clases del momento de estudio de la aplicación, primer encuentro con la tarea y construcción del discurso tecnológico. Las clases incluyeron a siete profesores e instituciones diferentes; cuatro de estos profesores realizaban dos tipos de clase; ordinaria y de preparación SIMCE. En el presente artículo se presentan dos de estas clases: momento de aplicación y otra del primer encuentro con la tarea.

Un componente esencial en el estudio son los análisis de las organizaciones matemáticas. Las OM fueron estudiadas en la evaluación SIMCE, en el programa de estudio y en las sesiones de clase. Los resultados obtenidos del estudio de las OM permitieron establecer conjeturas importantes sobre la influencia de la evaluación SIMCE en los diferentes niveles de co-determinaciones didácticos. Siendo ciertos resultados la base fundamental para el estudio de campo, a continuación se presenta brevemente dichos resultados.

4. Resultados preliminares al trabajo de campo: Análisis de las organizaciones matemáticas

El estudio de la evaluación SIMCE se realizó en un contexto comparativo considerando tres evaluaciones donde Chile participa: PISA, TIMSS y LLECE. El foco fue entender los motivos principales de esta evaluación, por lo que se estudió la visión de la enseñanza de las matemáticas a la que ella adhiere. Se analizó además las tareas específicas de la evaluación para comprender cómo esta visión se transpone en las tareas propuestas a los estudiantes y cuáles son las nociones geométricas que se destacan. Los principales resultados de este estudio mostraron que la visión de la evaluación SIMCE es definida de manera poco clara y que esto es en parte debido a la ausencia de un documento oficial que describa el instrumento de evaluación de forma extensa y precisa. Esto dificultaría el proceso de retroalimentación, puesto que el docente no cuenta con la información necesaria para comprender que saben o no sus estudiantes. Una concentración fuerte, fue constatada, sobre la evaluación del currículo, poniendo el acento sobre los conocimientos matemáticos más que competencias matemáticas. Del análisis de las tareas se concluyó que los tipos de tareas son principalmente de selección múltiple con una sola respuesta correcta. En su mayoría las tareas son presentadas en un contexto interno a las matemáticas con una predominancia de tareas de "reconocimiento y aplicación de conceptos y propiedades matemáticas" y de "razonamiento en resolución de problemas rutinarios" en segundo lugar.

Dado que la evaluación SIMCE es una evaluación que mide los contenidos del currículo nacional, los resultados del análisis de las organizaciones matemáticas del programa de estudio oficial chileno fueron utilizados comparativamente. Los principales resultados obtenidos mostraron que la visión de las matemáticas es más explícita que aquella de SIMCE. En términos generales el programa espera que los conocimientos que construyen los estudiantes les sean útiles a lo largo de la vida y que les permitan enfrentar los desafíos de la sociedad. La forma como el programa plasma esta visión es primeramente presentando tareas que tengan sentido para los estudiantes. Se cree que por este motivo se proponen tareas exploratorias, donde el alumno no solo aplica un conocimiento formal sino que además utiliza su experiencia de vida. Se verificó a través de las tareas el interés por llevar a los estudiantes a trabajar en diferentes contextos de la vida real, mediante la propuesta de una variedad de tareas. En contraste con las tareas

de la evaluación SIMCE el programa da importancia a los contextos externos a las matemáticas y al desarrollo de competencias. Al momento de analizar las organizaciones didácticas, el énfasis está en que los estudiantes sean capaces de comunicar, argumentar, razonar, conjeturar, justificar, tanto procedimientos como resultados. Incluso, a través de ciertas tareas se les hace contribuir bastante en el proceso de construcción del discurso tecnológico. Al finalizar el análisis de las OMPs y OMLs del programa, las tareas de la evaluación SIMCE representan solamente una pequeña parte de los tipos de tareas asociadas al momento de aplicación y evaluación propuestas por el programa de estudio. En consecuencia, la evaluación SIMCE se aleja de la visión de la enseñanza de las matemáticas que el programa espera transponer.

5. Estudio de campo: instituciones educativas y observaciones de clases

A través de las dos siguientes sesiones se propone mostrar los alcances de la evaluación SIMCE en el sistema educativo chileno. Los efectos que se han desencadenado en los últimos 10 años han llevado a una serie de cambios, incluyendo reformas curriculares, programas de mejoramiento de la educación, subvenciones especiales, extensión de las jornadas escolares y sobre todo la adopción de sistemas de evaluación estandarizados nacionales, como es la evaluación SIMCE. Hoy en día esta evaluación es reconocida dentro de la comunidad educacional como una herramienta fiable que mide la calidad de la educación nacional (García-Huidobro, 2002, p. 4). La validación de esta evaluación ha generado el deseo en las instituciones por obtener buenos resultados.

Por medio de este estudio de campo se quiso poner en evidencia el impacto directo de esta evaluación en diversas instituciones educativas, las cuales realizan múltiples acciones para mejorar sus resultados SIMCE. De igual forma se observan sesiones de clase con el objetivo de constatar como las prácticas de enseñanza podrías ser influencias por la evaluación SIMCE. Ambos estudios fueron enriquecidos con los datos obtenidos de las entrevistas y cuestionarios a los profesores.

5.1. Análisis de las instituciones educativas

El estudio se realizó en 12 establecimientos educacionales (públicos y semi-privados) de diferentes contextos socio-económicos en Santiago de Chile. Participaron 13 docente de matemática de octavo año de esos mismo establecimientos. Las herramientas utilizadas en la recolección de datos fueron ficha de los establecimientos, construidas con datos extraídos del Ministerio de Educación chileno, de cuestionarios y entrevistas realizadas a los profesores de matemáticas y al personal pedagógico (Directores y jefes de departamento de matemáticas). Para el estudio de las instituciones se consideraron tres dimensiones de análisis: el contexto socio-económico; el histórico de los resultados de la evaluación SIMCE y los dispositivos SIMCE utilizados. Este análisis permitió establecer categorías institucionales en relación a la evaluación SIMCE y mostrar el impacto directo de esta evaluación en las instituciones educativas.

5.2. Contexto de los establecimientos y dimensiones de análisis

Los datos recolectados de las instituciones educativas permitieron caracterizarlas y clasificarlas según grados de relación a la evaluación SIMCE. Al examinar los datos se establecieron diferentes criterios relacionados directamente a las instituciones, la

enseñanza y el desempeño de ellas en la evalaución SIMCE. A partir de los criterios que parecieron relevantes, se crearon dos dimensiones separadas utilizadas para caracterizar las instituciones: contexto socioeconómico[2] y las dos dimensiones institucionales de relación a SIMCE junto con los *criterios* que los constituyen en la tabla 1.

Tabla 1. Contexto socioeconómico y Dimensiones institucionales en relación a SIMCE

CARACTERÍSTICAS INSTITUCIONALES	CONTEXTO SOCIOECONÓMICO	DIMENSIONES DE RELACIÓN A SIMCE	
		1) Histórico SIMCE	2) Dispositivos SIMCE
Criterios	Grupo Socioeconómico Tamaño de clase Costo anual por alumno	Resultados de la evaluación SIMCE en 2004, 2007 y 2009 Los porcentajes de los niveles de logro de SIMCE 2009	Ensayo SIMCE, Talleres SIMCE Reforzamiento Contratación de personal externo

Fuente: Elaboración propia.

El contexto socioeconómico institucional caracteriza el entorno donde enseñan los profesores, mientras que cada una de las dimensiones institucionales caracteriza un aspecto de la relación de cada institución con SIMCE: Histórico SIMCE resume los resultados obtenidos por las instituciones en la evaluación SIMCE durante las tres últimas ediciones; Dispositivos SIMCE[3] describe los diferentes dispositivos puestos en marcha por las instituciones para mejorar los resultados SIMCE. Para caracterizar cada una de las instituciones y ver en que medida se han adaptado a enfrentarse a la evaluación SIMCE fue necesario explícitamente describirlas y compararlas según las dimensiones institucionales de relación con SIMCE. Las dos dimensiones están compuestas de varios criterios, expresados bajo la forma de indicadores cualitativos y/o cuantitativos. Por esto fue necesario transformar esta información en un sistema de datos cuantitativos, normalizados. A partir de estos resultados cuantitativos se puedo realizar comparaciones vectoriales entre los diferentes establecimientos. De este modo se determinaron los grupos de establecimientos que manifiestan similitudes vectoriales y así son clasificados según estas relaciones de proximidad.

[2] A través del estudio no se puede determinar si las características socioeconómicas de cada escuela tiene una relación directa con los resultados SIMCE obtenidos por cada institución o no. Por esto, en el análisis se consideran estas características como contexto externo a la dimensiones con relación a SIMCE

[3] Ensayo SIMCE: Una evaluación que integra tareas similares SIMCE y se realiza bajo condiciones similares.

Taller SIMCE: sesiones de clase extra-curriculares donde los estudiantes realizan tareas similares a las de la evaluación.

Reforzamiento: Las sesiones de clase extra-curriculares utilizadas para reforzar a los estudiantes con menos éxito académico.

Contratación de personal externo a la institución: Este personal es contratado para realizar diversas tareas con los estudiantes enfocadas a SIMCE. Los 2 últimos dispositivos se implementan en algunas instituciones para prepararse para el SIMCE, pero no son reconocidos oficialmente como tales.

5.2.1. Descripción de las categorías institucionales de relación a SIMCE

Basado en el análisis del nivel de relación a SIMCE por institución se definieron 3 categorías institucionales distintas. En la tabla 2 se presentan las características de cada categoría institucional, junto con las instituciones correspondientes.

Tabla 2. Categorías institucionales de relación a SIMCE

CATEGORÍA INSTITUCIONAL DE RELACIÓN SIMCE	**CAT.1: ALTA ACCIÓN – ALTO DESEMPEÑO**	**CAT. 2: ALTA ACCIÓN – MEDIANO DESEMPEÑO**	**CAT. 3: ALTA ACCIÓN –BAJO DESEMPEÑO**
Descripción	Instituciones con altos resultados en las evaluaciones SIMCE Varios dispositivos SIMCE usados Contexto socioeconómico favorable a las oportunidades escolares Colegios Particulares-subvencionadas (Costo compartido los con padres)	Instituciones con medianos resultados en las evaluaciones SIMCE Varios dispositivos SIMCE usados Contexto socioeconómico moderadamente favorable a las oportunidades escolares Escuelas Particulares-subvencionadas y Escuelas municipales	Instituciones con bajos resultados en las evaluaciones SIMCE Varios dispositivos SIMCE usados Contexto socioeconómico desfavorable a las oportunidades escolares Escuelas municipales (gratuitas)
Instituciones	Colegio Venezuela Colegio Colombia Escuela Brasil	Escuela Perú Colegio Argentina Colegio Chile Liceo Bolivia	Escuela Uruguay Escuela Paraguay Escuela Panamá Escuela Ecuador Escuela México

Fuente: Elaboración propia.

Se constató que a pesar de los diversos contextos socioeconómicos que existen, incluyendo las grandes diferencias en el costo por estudiante, el que varía de 0 a 380,000 pesos chilenos según el tipo de escuela, se han puesto en marcha varios dispositivos a través todas las instituciones educativas con la intención de mejorar los resultados SIMCE. Además, se descubrió que en la mayoría de los casos los dispositivos SIMCE puestos en marcha son los mismos, con la diferencia de como están organizados dentro de la institución. Los talleres SIMCE ocupan una parte importante dentro de las actividades escolares durante todo el año escolar. Consecuencia de esto, se constató una reducción de la enseñanza, donde se dejan de lado disciplinas que no serán evaluadas (ej. música, religión, inglés) durante periodos más o menos largos par enfocarse en las que sí serán evaluadas. Tal constatación fue mencionada por los profesores durante las entrevistas. Los profesores también manifestaron la disminución del tiempo que tienen para abarcar los contenidos del currículo; lo que en algunos profesores es una fuente de estrés y rechazo a la evaluación SIMCE. Se evidenció una situación particular con algunos estudiantes donde el reforzamiento se realiza en las asignaturas antes mencionadas, por lo que ellos no participan en la totalidad de las disciplinas definidas en el currículo oficial. La reducción curricular en la disciplina de matemáticas también fue observada dado que los profesores declaran poner énfasis en los contenidos más evaluados por SIMCE. Dentro del personal externo contratado se observó gente con diferentes formaciones (ej. psicopedagogos, ingenieros, estudiantes de matemáticas, universidades y organizaciones privadas) que brindan servicios adicionales dentro la

institución: proveedores de material, organismos evaluadores simulando la evaluación SIMCE y entregando estadísticas por alumnos y seguimiento permanente. En las entrevistas algunos profesores manifestaron "desaprobación y pocos beneficios aportados por el personal externo en el aprendizajes de los estudiantes" (Ruminot Vergara, 2014, pp. 352-354).

Para profundizar aún más sobre la influencia de la evaluación SIMCE sobre las prácticas docentes, se presentan los principales resultados de estudio de las sesiones de clase ordinarias.

5.3. Análisis de las observaciones de clases

El estudio de las prácticas de enseñanza es una tarea compleja. Las acciones de los profesores se ven influenciadas por diversas restricciones y limitaciones que condicionan sus actividades pedagógicas. Para determinar si las acciones de profesor son influenciadas o no por SIMCE, se creó una metodología para el análisis de las sesiones de clase, haciendo intervenir de manera comparativa las categorías de la gestión pedagógica del profesor durante las sesiones de clase ordinaria y las destinadas a SIMCE.

Las observaciones de clase se hicieron en 9 diferentes instituciones de la región metropolitana de Santiago de Chile. Se observaron en total 17 sesiones de 8° año - clases ordinarias y talleres SIMCE - realizados por 9 docentes. Para este estudio se seleccionaron dos profesores por ser representativos de las prácticas de enseñanza observadas.

Se hizo en primer lugar, un trabajo de redacción de sesiones de clase que permitieron construir fichas para las sesiones de clase a partir de dos dimensiones. Una dimensión que consideró las organizaciones matemáticas propuestas por el profesor y otra dimensión orientada hacia las organizaciones didácticas puesta en marcha. En el análisis de las sesiones de clase también se incorporó el perfil de los docentes y de las categorías institucionales definidas en el trabajo de tesis, que permitieron caracterizar y obtener resultados sobre la influencia de la evaluación SIMCE en la elección de las tareas y su gestión didáctica. En la elaboración de la ficha para analizar las transcripciones de clase se consideraron: i) el contexto del establecimiento, ii) el contexto de clase, y iii) la gestión didáctica de la enseñanza.

5.3.1. Análisis de datos - Construcción de fichas de clase

A partir de las narraciones de sesiones de clase se identifican regularidades entre ellas. Para poner en evidencia aquellas características se construyó un modelo de ficha de las observaciones por clase. Cada ficha de clase incluyó una descripción del contexto del establecimiento dentro del cual se realizaron las observaciones. Esto permitió considerar el contexto y las co-determinaciones didácticas que pesan sobre el profesor. También se consideraron las características profesionales y la experiencia del docente. De igual forma se tuvo en cuenta el contexto de cada clase examinada, que incluyó: el contenido matemático y los tipos de tareas estudiadas durante la sesión; el tipo de clase, el momento de estudio en desarrollo y los recursos pedagógicos utilizados por el profesor y por los estudiantes. Finalmente, se distinguieron las interacciones entre profesor y estudiante según su naturaleza, en formas de interacción individuales y colectivas.

5.3.2. Ficha de observaciones de sesiones de clase ordinarias

a) Ficha 1. Profesor Dunas del establecimiento Colegio Argentina

Contexto del establecimiento

El establecimiento de está en la Categoría 2: Alta acción y medianos resultados. El contexto socio económico es parcialmente favorable a las oportunidades de aprendizaje. El perfil del profesor se sitúo en los niveles que van de intermedio-bajo a intermedio alto en las cuatro dimensiones (Formación y experiencia, Posicionamiento institucional, Relación a la evaluación SIMCE, Sensibilidad Didáctica). Dentro de su establecimiento el profesor Dunas no realiza ningún dispositivo, solamente tiene la responsabilidad de reforzar los contenidos que arrojan bajos resultados en los ensayos SIMCE.

Contexto de la clase

La clase que observó fue sobre el contenido de proporcionalidad y porcentajes. Ella corresponde al momento del estudio de aplicación. El profesor utiliza principalmente la pizarra como recurso de aula, mientras que los estudiantes utilizan sus cuadernos y una guía de trabajo. El profesor Dunas propuso dos tipos de tareas utilizando principalmente una misma técnica de resolución. Vemos claramente la especificación del contrato didáctico en cada una (Brousseau, 1988).

El primer tipo de tarea (figura 1) que los estudiantes debieron resolver fue el cálculo de un porcentaje dado de una suma de datos (figura 2):

"Calcula el 10% de 184.000; 12% de 155.003; 27% de 600.000"

Figura 1.Tarea de la clase de Profesor Dunas
Fuente: Elaboración propia.

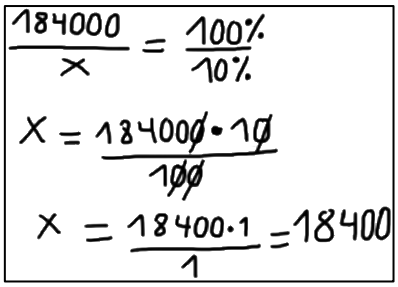


Figura 2. Técnica de resolución de tarea figura 1 del Profesor Dunas
Fuente: Elaboración propia.

Para este tipo de tareas el profesor explicitó la técnica siguiente (figura 3):

Catalina compró un par de botas a 14.300 pesos, en la liquidación de una tienda. Si todo el calzado estaba rebajado en un 35%. ¿Cuál era el precio original de las botas? Original de las botas quiere decir sin el descuento.

Figura 3. Situación problema del Profesor Dunas
Fuente: Elaboración propia.

Para este tipo de tarea también se puede proponer una técnica más directa, el x% de una magnitud se obtiene al multiplicarla por $\frac{x}{100}$. La técnica propuesta por el profesor es más próxima al cálculo de una cuarta proporcional, apoyándose en la técnica de producto cruzado.

Otro tipo de tareas (figura 3) que el profesor propuso corresponde al cálculo de un precio inicial conociendo el precio final y el porcentaje de descuento. El profesor es quien explicitó una técnica de presentación gráfica para su resolución.

Este tipo de tarea es más complejo que lo anterior, como lo muestra la literatura. La técnica que utiliza el profesor consiste en determinar el coeficiente multiplicador, entre el precio inicial y final, que en este caso es $\left(1-\frac{35}{100}\right)$=0,65, luego apoyándose de la misma expresión trabajada resolver la ecuación que permite encontrar el valor de la variable precio original. Consideramos que esta técnica gráfica es necesario complementarla con una resolución mediante expresiones algebraicas que permitan hacer referencia a las propiedades relativas a la disminución o al aumento de una magnitud.

Gestión didáctica

En las observaciones de clases del profesor Dunas se constató que el desarrollo de la clase es cíclico, con formas de trabajo colectivas e individuales. Retomando la tarea (figura 4) para analizar cómo a partir de la representación gráfica el profesor la explicó a los estudiantes. Se seleccionó esta tarea porque fue la primera de este tipo y con ella el profesor explicita la técnica.

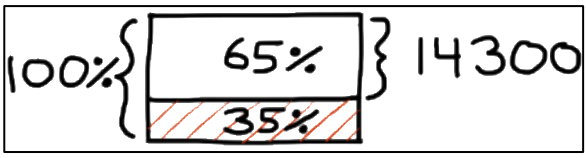


Figura 4. Técnica de resolución de la situación problema
Fuente: Elaboración propia.

El profesor presentó la tarea de la figura 5 con el objetivo de anticipar los posibles errores de interpretación que pudiesen realizar los estudiantes. A continuación, se ilustra a través de una fase los argumentos utilizados por el profesor para explicar la tarea y la técnica. El profesor explicó la tarea apoyándose en una técnica gráfica y centrando su atención en establecer una relación correcta entre las variables:

> *1. Prof.: Quiero que hagamos el siguiente ejercicio primero. Cierre el cuaderno y mire a la pizarra. 'Catalina fue a comprar las botas y estaban con un 35% de descuento'. Aquí es donde los alumnos se equivocan y piensas que lo que pagó Catalina es el 35 % de las botas, eso está mal. Los alumnos dicen ella pagó 14.300 pesos y eso es el 35 % o piensan que eso es el 100%. Eso es un error. El profesor realiza el esquema de la Figura 4 en la pizarra.*
>
> *2. Prof.: El total de las botas es el 100%, el precio completo esta ahí. A ese 100% le descontaron el 35%, le quitaron esa cantidad. Y eso que queda acá es lo que pagó Catalina, el 14.300 pesos, ¿y que porcentaje es? El 65%, ella pagó solamente el 65% ¿Cómo voy a armar la proporción? ¿Tengo el total de las botas?*
>
> *3. Alums.: No.*
>
> *4. Prof.: entonces, x = 100% ¿Cuál es la cantidad parcial que tengo? 14 300 = 65%*

5. Prof.: Ya tengo la proporción. Ojo, los errores que comenten los estudiantes, es pensar que 14.300 pesos es el precio total o es el 35%. El problema dice que le rebajaron el 35% al precio original de las botas. Con esta proporción puede calcular el precio de las botas. Preguntas que pueden aparecer: ¿Cuál es el precio original de las botas?, ¿cuál fue el descuento en plata que le hicieron? Sabemos que al precio total le descontaron el 35%, pero no sabemos el precio original. Resuélvalo.

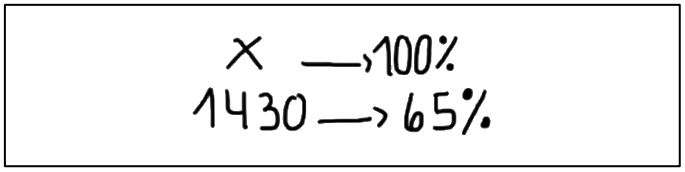


Figura 5. Síntesis de la técnica -Prof. Dunas
Fuente: Elaboración propia.

El profesor es consciente de las dificultades de la tarea y de los errores asociados. Él pone en obra una esquematización gráfica y un discurso que llevará a posicionar los números, siendo uno de ellos la incógnita, para escribir la proporción. No obstante, las propiedades ligadas a este tipo de tarea no son explotadas, en particular la aplicación de la variación de un porcentaje no es abordado, siendo que este tipo de tarea permitiría abordar esta propiedad, el trabajo es focalizado en la identificación de la cantidad parcial y su porcentaje. Mediante el desarrollo de la técnica subrayamos que la reducción de la magnitud no se explica, que la identificación entre magnitud y cantidad y la operatoria que existe entre ellas no se pone en relieve. Por ejemplo, al multiplicar una magnitud por 0,65 permite explicar una variación de - 35% $porque$ $(0{,}65 - 1)x100 = -35$. Como se señala precedentemente el coeficiente multiplicador de la magnitud no es trabajado.

El profesor presentó tipos de tareas con pequeñas variaciones que hacen trabajar una misma técnica. En esta fase además se observó que se refuerza la técnica de la simplificación de fracciones. La participación de los estudiantes es permanente, pero de forma verbal, dado que el profesor es el que realizó las correcciones en la pizarra.

b) Ficha 2. Profesor Ocaña del establecimiento Liceo Bolivia

Contexto del establecimiento

La institución está en la Categoría 2, como la del profesor Dunas descrito anteriormente. Dentro de la categoría de perfil docente el profesor tuvo el nivel más bajo de todos en las dimensiones formación y experiencia y posicionamiento institucional, pero se destacó por tener la más alta relación con la evaluación SIMCE en gran parte esto fue por su implicación en varios niveles. Dentro de su establecimiento el profesor Ocaña realiza tres dispositivos SIMCE, el Taller SIMCE, el Ensayo SIMCE y Reforzamientos.

Contexto de clase

La sesión corresponde a la unidad de geometría. La clase que se observó fue sobre la circunferencia, perímetro y área. Ella corresponde al momento del estudio del primer encuentro con el tema, donde se dan a conocer la definición de circunferencia, sus propiedades, el perímetro y el área. El profesor solo utilizó la pizarra para desarrollar la clase, mientras que los estudiantes utilizaron sus cuadernos. Las nociones matemáticas presentadas fueron "*la circunferencia y sus elementos básicos y cómo calcular área y perímetro*". El profesor escribió los títulos y subtítulos de los contenidos en la pizarra y luego dictó las definiciones a los estudiantes para que las copien en sus cuadernos. Esto fue

complementado con un diseño, en este caso una circunferencia y sus elementos como el diámetro, radio, cuerda, arco, tangente y secante de la circunferencia. Además una explicación de las definiciones.

Gestión didáctica

La organización didáctica se realizó en dos fases. Una forma colectiva débilmente interactiva y otra forma individual de corta duración. En la forma colectiva el profesor organizó su clase a través del dictado del contenido. Él no les hizo preguntas a los estudiantes, ni ellos tampoco realizaron preguntas. La forma individual de trabajo es corta y se limita a copiar lo que el profesor escribe en la pizarra: diseño de la circunferencia y ejemplo de una tarea sobre el perímetro del círculo.

La clase del profesor Ocaña se inscribe en un modelo de clase tradicional, donde comienza el estudio por un momento de institucionalización, entregando las definiciones y propiedades de las nociones a desarrollar y luego se realizan tareas de aplicación de la técnica. Las definiciones fueron del siguiente tipo:

> *Prof.: La circunferencia: es el conjunto de puntos que equidistan (tienen la misma distancia) de un punto fijo, llamado centro de la circunferencia. La distancia que existe entre el centro de la circunferencia y cualquier punto de ella, se denomina radio. El radio nos va a servir después para calcular área y perímetro.*

El profesor retomó esta definición y la explicó apoyándose de un diseño que realizó en la pizarra, poniendo énfasis en la circunferencia como el conjunto de puntos que están a igual distancia del centro. De manera general el realizó el mismo trabajo con cada definición.

El profesor Ocaña no precisa que las definiciones de recta tangente y recta secante son particulares a la circunferencia. Una vez que termina las explicaciones vuelve a copiar un título: "Área y perímetro de la circunferencia" e introduce una frase recordatoria (línea nº 2) sobre cómo calcular el perímetro de un cuadrado y de un triángulo.

> *1. Prof.: Cuando calculábamos el perímetro de una figura geométrica lo que hacíamos era sumar la medida de todos los lados. Cuando sacábamos el perímetro del cuadrado era la suma de los cuatro lados; del triángulo la medida de los tres lados. Entonces, pongan ahí, para calcular el perímetro de una circunferencia ocuparemos la siguiente fórmula, ponga ahí bajo* $P = 2\pi r$ *¿Qué pasa cuándo no tengo nada aquí? ¿Qué hay ahí?*
>
> *2. Alum. 1: Un uno,*
>
> *Prof.: No, hay una multiplicación, entonces multiplico 2 por* π *por r. Esa fórmula es la que vamos a utilizar para calcular el perímetro. Esta, chiquillos, se la tienen que aprender de memoria. Ya copin el siguiente ejemplo* (figura 6).

"Determinar el perímetro de la circunferencia de radio 5cm ($\pi = 3{,}14$)"

$$P = 2 \times 3{,}14 \times 5$$

$$P = 10 \times 3{,}14$$

$$P = 31{,}4$$

Figura 6. Tarea de la clase de Profesor Ocaña
Fuente: Elaboración propia.

Se observa a partir de la línea nº 4 el profesor presenta el perímetro del círculo sin dar ninguna explicación sobre la naturaleza de la medición de esta magnitud. Se introduce la noción de perímetro como la suma de los lados del contorno de una figura poligonal y pasa inmediatamente a la fórmula del perímetro, sin más explicaciones. Se nota una gran diferencia con las sugerencias hechas en el programa de estudio para determinar las fórmulas del perímetro y del área de la circunferencia enfatizando la proporcionalidad entre el diámetro y perímetro y la naturaleza del número pi (π). El profesor entrega la fórmula, sin explicar que significa ni cómo se puede obtener, solamente señala que pi (π) es una constante que deben reemplazar por el número 3,14.

> *Prof.: ¿Qué tengo que hacer aquí? Tengo que determinar el P ¿cuánto mide la circunferencia en todo su contorno? El radio del la circunferencia mide 5 cm, el perímetro va a ser 2 (□) que es una constante que vamos a utilizar (apunta el enunciado donde escribió 3,14) por el radio. Entonces, lo que vamos hacer es reemplazar los datos. ¿Ya cómo lo reemplazo? Dígame Pablo. (A medida que el alumno responde el profesor va resolviendo el ejercicio)*
>
> *Alum.1: 2 □ 3,14 □ 5*
>
> *Prof.: Cinco veces dos 10 por 3,14 ¿Qué se hace con la coma? La traslado*
>
> *Alums.: La traslado*
>
> *Prof.: ¿Para dónde? Para allá (señala con la mano hacia derecha) ¿Cuánto queda? 31,4 cm. Acuérdense de esto, cuando multiplicamos por 10 estamos amplificando. ¿Qué hago con la coma? ¿Si hubiera multiplicado por 100 que pasa con la coma?*
>
> *Alums.: Corro dos lugares.*
>
> *Prof.: ¿Alguien no entendió?*
>
> *Alum.2: Yo*
>
> *Prof.: Ok, ponga atención. ¿Cuál es el perímetro de una circunferencia de radio de 5cm? Dijimos que el perímetro va a ser esto (el profesor muestra la fórmula). Siempre va a ser esto.*
>
> *Alum.2: ¿Siempre va a ser esa?*
>
> *Prof.: Siempre. Esta es la fórmula, el 2 se mantiene, pi dijimos que valía 3,14*
>
> *Alum.2: ¿Siempre va valer 3,14?*
>
> *Prof.: Sí. Ya y después es esto (muestra la parte final de la resolución)*

La duda de la alumna permite poner en evidencia que la naturaleza de la fórmula y su utilización no son para nada evidentes para los estudiantes. Sin embargo, el profesor pasa por alto cualquier explicación sobre la fórmula que permita a los estudiantes comprender mejor la medición de la circunferencia. Los estudiantes asumen el contrato didáctico que impone el profesor de aprenderse la fórmula de memoria, ellos aceptan que deben colocar en esa fórmula los datos que les permitirán determinar el perímetro de una circunferencia. Como se señala anteriormente a diferencia de lo que plantea el programa de estudio el profesor no retoma ninguna de las actividades propuestas para trabajar la noción de perímetro de la circunferencia, ni para poner en evidencia que el número pi (π) es la constante de proporcionalidad entre el diámetro y el perímetro.

5.3.3. Resultados de las observaciones de clase

A través de estas observaciones se identificó tipos de clases que se aproximan a momentos de estudio específicos: *el momento del primer encuentro con un tema, el momento de construcción del discurso tecnológico* y *los momentos de aplicación.* El momento del primer

encuentro con un tema se caracteriza de modo predominante por una fase de institucionalización y luego de aplicación de la técnica previamente introducida. Las organizaciones matemáticas y las organizaciones didácticas sugeridas por el programa de estudio, son poco presentes. En particular no se encontraron tareas de exploración, ni técnicas que emerjan del uso de material concreto. Se observó, también, la ausencia de un trabajo que permita la articulación entre la tarea y la técnica para construir un discurso tecnológico (por ejemplo, el caso de la circunferencia).

En las sesiones de clases correspondientes al momento de aplicación se observó que se privilegia un cierto tipo de tarea y se pone el énfasis en el trabajo de la técnica, especialmente de una sola técnica. En la elección del tipo de tarea no se observan los tipos de tareas que propone el programa de estudio para estos momentos, como por ejemplo: tareas de demostración, tareas de construcciones geométricas, tareas de cálculo directo y tareas de problemas rutinarios. En estas sesiones de clase fueron trabajadas principalmente tareas de cálculo directo y problemas rutinarios. En consecuencia, se constató un distanciamiento entre las organizaciones matemáticas que propone el programa para estos momentos del estudio con las organizaciones matemáticas seleccionadas y realizadas en estas sesiones de clase. Además, se evidenció que las tareas presentadas son muy similares a las tareas de talleres y ensayos SIMCE.

Desde el punto de vista de la gestión didáctica las *formas de interacción colectiva* fueron más presentes en las sesiones de clase correspondiente al momento de aplicación, un poco menos presentes en el momento de construcción del discurso tecnológico y casi ausentes en las sesiones de clase del primer encuentro con un tema. Por medio de estas interacciones se constató que por medio de un trabajo de exploración de situaciones e intercambio preguntas y respuestas se trata de dar sentido a los contenidos. También, se enfatizan los errores, constituyendo una fuente de razonamientos y co-evaluación.

6. Discusión y conclusiones

Se constata que la evaluación SIMCE refleja muy parcialmente la visión de la enseñanza de las matemáticas del programa de estudio. Los resultados del análisis de las tareas accesibles de la evaluación SIMCE muestran en efecto que estas tareas son principalmente situadas en un contexto interno a las matemáticas y de aplicación directa de conceptos y propiedades matemáticas. En contraste a las tareas tipo SIMCE, al analizar las organizaciones matemáticas puntuales del programa de estudio se encuentra una gran diversidad de tareas: de reconocimiento, de construcción con instrumentos geométricos, de aplicación y cálculo, de demostración, de construcción de fórmulas y la presencia frecuente de tareas contextualizadas. Examinando SIMCE respecto a estas categorías, solo se evidencia tareas de aplicación y cálculo, además de algunas tareas contextualizadas, pero rutinarias. Entendiendo que todas las tareas SIMCE no son públicas, al menos esto tiende a confirmar que la evaluación SIMCE representa parcialmente el programa nacional de estudio.

La utilización de los dispositivos identificados en el estudio afectan ampliamente las prácticas de enseñanza. En algunos casos la puesta en marcha de os dispositivos conducen a una contracción de ciertas disciplinas, como los son: Inglés, Religión, Música, Educación Física y Artes. Estas disciplinas no son impartidas durante un mes precedente de la evaluación SIMCE o en algunos casos durante un periodo más extenso.

En general, los horarios correspondientes a estas disciplinas son utilizados para realizar los ensayos SIMCE y/o poner en marcha guías de ejercitación y de reforzamiento. Igualmente se observa una contracción en el programa mismo de matemática. Por un lado, el trabajo de las unidades temáticas se suspende uno o dos meses antes de la evaluación, en agosto y septiembre, luego se retoma después de la evaluación. Esta reducción de tiempo de la enseñanza hace que no todos los contenidos se logren tratar con los estudiantes. Por otro lado, los profesores declararon no contar con suficiente tiempo para abarcar todos los contenidos que deben enseñar. Algunos profesores nos expresaron que ciertos contenidos los dejan para el final del año y los trabajan si les queda tiempo (Ruminot Vergara, 2014).

Continuando en esta dirección, para ver si había efectivamente reducciones alrededor de los contenidos y de los tipos de tareas evaluadas, se observan sesiones de las clases ordinarias. Las sesiones correspondientes al primer encuentro con el tema se inscriben en un modelo de enseñanza clásico, donde se comienza con un momento de institucionalización que es seguido por la ejercitación de la técnica. Esta característica igualmente se observa en la sesión de construcción del discurso tecnológico. Desde el punto de vista de las organizaciones matemáticas, el énfasis es puesto en la descripción de la técnica, pero se observan imprecisiones en la definición de propiedades en el momento de institucionalización. En las sesiones correspondientes al momento de aplicación, se trabaja con un tipo de tareas, principalmente de aplicación directa y problemas rutinarios. En consecuencia, las organizaciones matemáticas observadas reflejan de modo muy limitado las organizaciones matemáticas que sugiere el programa de estudio. El tipo de tareas exploratorias que le permitan a los estudiantes un encuentro con un tipo de tarea (en el sentido de los momentos del estudio) y el desarrollo de la técnica no están presentes de forma significativa. No se puede, a priori, poner en relación directa esta constatación como un efecto de la evaluación SIMCE sobre las prácticas docentes. Sin embargo, si consideramos que los profesores nos señalaron en la entrevista una contracción del currículo causada por el tiempo pasado en preparar la evaluación SIMCE, podríamos pensar que la ausencia de tareas que demandan mayor inversión de tiempo como las propuestas por el programa de estudio son menos trabajadas, y que se privilegian tareas de ejercitación de técnicas previamente introducidas después de una institucionalización precoz.

Se identifican efectos directos e indirectos de la evaluación SIMCE sobre las prácticas de enseñanza de matemáticas. Entre los efectos directos, aquellos que tiran la atención son naturalmente los dispositivos SIMCE. La investigación se centra en la caracterización de estos dispositivos, en la comprensión de su puesta en marcha y trata de focalizar sus efectos sobre las prácticas de enseñanza. Se puso en evidencia los efectos negativos de esos dispositivos como la contracción de disciplinas no evaluadas por SIMCE, y la disminución del tiempo escolar disponible para trabajar las unidades temáticas del currículo. Sin embargo, no se ha abordado la pregunta de la eficacia de estos dispositivos. Los resultados de la evaluación 2011 (comunicados en 2012) muestran una estabilidad o un ligero retroceso según la institución en relación a los resultados de los años 2004, 2007 y 2009 para las instituciones estudiadas. Estos resultados proponen claramente la pregunta de la eficacia real de los numerosos dispositivos puestos en marcha, y de los límites de una acción didáctica organizada alrededor de la preparación de una evaluación, aunque ella sea de calidad.

Referencias

Artigue, A., Coagri Nassouri, C., Smida, H. y Winslow, C. (2012). *Évaluations Internationales: Impacts politiques, curriculaires et place des pays francophones. Project spécial 2. Espace Mathématique Francophone.* Recuperado de goo.gl/fpcfZ5

Bodin, A. (enero, 2006). *Les mathématiques face aux évaluations nationales et internationales. De la première étude menée en 1960 aux études TIMSS et PISA... en passant par les études de la DEP et d'EVAPM.* Comunicación presentada en el Séminaire de l'EHESS, Institut de Recherche sur l'Enseignement des Mathématiques (IREM) de Franche-Comté.

Bodin, A. (2007). Dissonances et convergences évaluatives - De l'évaluation dans la classe aux évaluations internationales: Quelle cohérence? *Bulletin de l'APMEP, 474*, 47-79.

Bosch, M. y Gascón, J. (2003). Les praxéologies didactiques. Cours 2 - Théories & Empiries. En J. L. Dorier, M. Artaud, M. Artigue, R. Berthelot y R. Floris (Eds.), *Actes de la 11e École d'été de didactique des mathématiques* (pp. 23-40). Grenoble: La Pensée Sauvage.

Brousseau, G. (1988). Le contrat didactique: Le milieu. *Recherches en didactique de mathématiques, 19*(3), 303-336.

Castela, C., Consigliere, L., Guzman, I., Houdment, C., Kuzniaz, A., Rauscher, J-C. (2006). *Paradigmes géométriques et géométrie enseignée au Chili et en France. Une étude comparative de l'enseignement de la géométrie dans les systèmes scolaires chilien et français.* París: IREM.

Chevallard, Y. (2002). Organiser l'étude 1. Structures et Fonctions. En J. L. Dorier (Ed.), *Actes de la 11e Ecole d'été de didactique des mathématiques -Corps- 21-30 Août 2001* (pp. 3-22). Grenoble: La Pensée Sauvage.

Chevallard, Y. (1999). La recherche en didactique et la formation des professeurs: Problématiques, concepts, problèmes. En J. L. Dorier (Ed.), *Actes de la X Ecole d'été de Didactique* (pp. 98-112). París: Académie de Caen.

Chevallard, Y. (2002). Les praxéologies didactiques. Cours 3 - Ecologie & Régulation. En J. L. Dorier, M. Artaud, M. Artigue, R. Berthelot y R. Floris (Eds.), *Actes de la 11e École d'été de didactique des mathématiques, Corps (Isère), du 21 au 30 août 2001* (pp. 41-56). Grenoble: La Pensée Sauvage.

Clarke, D. (2003). International comparative research in mathematics education. En A. J. Bishop, M. A. Clements, C. Kei- tel, J. Kilpatrick y F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 143–184). Dordrecht: Kluwer Academic.

Cox, C. (2003). *Políticas educacionales en el cambio de siglo: La reforma del sistema escolar de Chile.* Santiago de Chile: Editorial Universitaria.

García-Huidobro, J. E. (2002). *Usos y abusos del SIMCE.* Santiago de Chile: Universidad Alberto Hurtado.

LLECE. (2008). *Los aprendizajes de los estudiantes de América Latina y el Caribe. Primer reporte de los resultados del Segundo Estudio Regional Comparativo y Explicativo.* Santiago de Chile: OREALC/UNESCO.

LLECE. (2009). *Aportes para la enseñanza de la Matemática. Segundo Estudio Regional Comparativo y Explicativo.* Santiago de Chile: OREALC/UNESCO.

MINEDUC. (2007). *Mapas de progreso – Números, operatoria, geometría, algebra, datos y azar. Unidad de currículo.* Santiago de Chile: Autor.

MINEDUC. (2010). *Niveles de logros. Para 8avo. Año básico de Matemáticas – SIMCE. Unidad de currículo y evaluación.* Santiago de Chile: Autor.

MINEDUC. (2010). *Orientaciones para Docentes Educación Básica - SIMCE 2013*. Santiago de Chile: Autor.

Mons, N. (2009). *Les effets théoriques et réels de l'évaluation standardisée. Les évaluations standardisées des élèves en Europe: Objectifs, organisation et utilisation des résultats EACEA*. Bruselas: Eurydice.

OCDE. (2003). *Assement framework - PISA 2003. Mathematic, Reading, Science and Problem Solving, Knowledge and Skills*. París: Autor.

OCDE. (2004). *Revisión de políticas nacionales de educación*. París: Autor.

OCDE. (2009). *Assessment framework – PISA 2009. Key Competencies in reading, mathematics and science*. París: Autor.

OCDE. (2009). *Take Thes Test – PISA. Samples questions from OECD's PISA assessments*. París: Autor.

OCDE. (2013). *Cadre d'évaluation et d'analyse du cycle - PISA 2012. Compétences en mathématiques, en compréhension de l'écrit, en sciences, en résolution de problèmes et en matières financières*. París: Autor.

OCDE. (2006). *Compétences en sciences, lecture et mathématiques: Le cadre d'évaluation*. París: Autor.

Ministerio de Educación. (2002). *Programa de Estudio. Octavo año escolar. Unidad de currículo y evaluación*. Santiago de Chile: Autor.

Ruminot Vergara, C. (2009). *SIMCE: Analyses à niveau micro et macro institutionnelle - Outil méthodologique d'analyses, pour déterminer l'incidence des relations institutionnelles en la qualité de l'éducation chilienne* (Trabajo fin de master). Universidad de Paris-Diderot, París.

Ruminot Vergara, C. (2014). *Los efectos de un sistema nacional de evaluación estandarizado sobre las prácticas de enseñanza de las matemáticas: Caso de SIMCE en Chile* (Tesis doctoral). Universidad de Paris-Diderot, París.

Schoenfeld, A. (2007). *Assessing Mathematical Proficiency*. Cambridge: Cambridge University Press.

TIMSS. (2011). *Assessment frameworks of TIMSS & PIRLS*. Recuperado de http://timssandpirls.bc.edu/

Breve CV de la autora

Carolina Ruminot Vergara

Carolina actualmente forma a futuros professores en la Universidad de Ottawa, Canadá, en temas que incluyen la didáctica de las matemáticas, la evaluación y las teorías actuales de enseñanza y aprendizaje. Obtuvo su doctorado en Didáctica de las Matemáticas en 2014 de la Universidad Paris Diderot, Francia. Su investigación de tesis se centró en los efectos de un sistema de evaluación estandarizado sobre las prácticas de enseñanza de las matemáticas, tomando como caso particular SIMCE (Sistema de Medición de la Calidad de la Educación) en Chile. Ella se interessa en la vinculación de sus conocimientos de la didáctica francesa con las investigaciones etnológicas y sociales en el campo de las matemáticas. Está explorando vías para hacer las matemáticas más accesibles a estudiantes de diferentes contextos culturales y étnicos, muy presentes en Canadá, y continua estudiando las limitaciones de los sistemas de evaluaciones estandarizadas sobre los sistemas de enseñanza en contextos multiculturales. Código ORCID: 0000-0001-8597-1102. Email: caruminot@gmail.com

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 89-107.
https://doi.org/10.15366/riee2017.10.1.005

Creación, Desarrollo y Resultados de la Aplicación de Pruebas de Evaluación basadas en Estándares para Diagnosticar Competencias en Matemática y Lectura al Ingreso a la Universidad

Creation, Development and Assessment Tests Standards-based Application Results to Diagnose Math and Reading Skills to University Entrance

Pilar Rodríguez Morales *

Universidad de la República

El artículo presenta el proceso de creación y desarrollo de pruebas de evaluación de las competencias en Matemática y Lectura para el nivel de ingreso a la Universidad y su aplicación a los ingresantes a los Centros Universitarios Regionales de la Universidad de la República (Uruguay) que atraen a un perfil de estudiantes con vulnerabilidad social y académica. La metodología utilizada para el desarrollo de las pruebas implicó la creación y establecimiento de estándares de contenido a través de grupos de expertos. La clasificación en los estándares de desempeño se obtuvo a través de un método nuevo propuesto por García et al. (2013) basado en la Teoría de Respuesta al Ítem (TRI). Se obtuvieron dos pruebas unidimensionales, con muy buena consistencia interna, validadas y calibradas mediante TRI que clasificaron a los estudiantes en tres niveles de desempeño (insuficiente, suficiente y avanzado). El 22% superó la suficiencia en la prueba de Matemática y el 53% lo logró en la prueba de Lectura. Aunque la proporción de estudiantes que alcanzaron la suficiencia es baja, estos resultados son coherentes con otras investigaciones e impulsaron líneas de acción que incluyeron la creación de grupos académicos en Lectura y Matemática para el diseño de programas de apoyo.

Palabras clave: Pruebas diagnósticas, Prueba de matemática, Prueba de lectura, Ingreso a la universidad, Estándares académicos.

The article presents the process of creation and development of assessment tests in Math and Reading skills for entry level to university and its application to entrants to Regional University Centers of the University of the Republic (Uruguay) that attract a profile of students with social and academic vulnerability . Methodology used for tests development involved creation and establishment of content standards through expert groups. Performance standards' classification was obtained through a new method proposed by Garcia et al. (2013) based on the Item Response Theory (IRT). Two one-dimensional tests were obtained, with very good internal consistency, validated and calibrated by IRT that classified students in three performance levels (insufficient, sufficient and advanced). The 22% overcame the sufficiency in Math test and 53% it reached in Reading test. Although students' proportion of achieving proficiency is low, these results are consistent with other research and it promoted action lines that included the creation of academic groups in Reading and Mathematics for the design of support programs.

Keywords: Diagnostic tests, Mathematics tests, Reading tests, College admission, Academic standards.

*Contacto: prodriguez@cure.edu.uy

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 26 de octubre de 2016
1ª Evaluación: 2 de diciembre de 2016
Aceptado: 22 de diciembre de 2016

1. Introducción[1]

El ingreso a la Educación Superior supone una serie de retos que los estudiantes deben enfrentar, siendo la exigencia académica uno de los que plantea mayores dificultades. Por este motivo, el desempeño de los estudiantes de ingreso a la universidad se ha convertido en objeto de diversas investigaciones (Attewell, Heil y Reisel, 2012).

Los sistemas universitarios latinoamericanos tienen procesos de ingreso muy disímiles entre sí, encontrándose algunos sistemas con una selección muy competitiva y estricta hasta otros con acceso libre e irrestricto como es el caso del Uruguay. La Universidad de la República de Uruguay (Udelar), la principal institución universitaria del país, habilita el ingreso a todos los que acrediten haber finalizado el nivel educativo previo. Por otro lado, la educación media superior (bachillerato) tampoco realiza pruebas de evaluación sistemáticas sobre los aprendizajes de sus egresados, la última se realizó en 2003. Se torna necesario, entonces, la evaluación diagnóstica de las competencias al ingreso a la Universidad, ya que acceden a la educación superior estudiantes con diferentes niveles académicos. La información proporcionada por este tipo de evaluaciones es de gran utilidad para la planificación de acciones de mejora y la orientación de las trayectorias académicas de los estudiantes.

La Educación Superior de Uruguay no cuenta con una extensa tradición en realizar pruebas estandarizadas, como existe en otros niveles del sistema educativo, principalmente Primaria. Sin embargo, la Udelar ha desarrollado pruebas de evaluación de tipo diagnóstico para los estudiantes de ingreso desde las distintas facultades (Altmark et al., 2006; Enrich et al., 2006; Míguez, Blasina y Loureiro, 2013; Mussio y Martinotti, 2013; UEFI, 2012; Unidad de Enseñanza de Facultad de Ciencias, 2005, 2010). El principal objetivo de estas evaluaciones es conocer las competencias de los estudiantes que ingresan en las áreas consideradas básicas para la carrera en que se matricularon (matemática, física, química y comprensión lectora) con el fin de facilitar la transición entre ciclos educativos. No obstante, todas estas experiencias se han circunscripto a una facultad o área, no se basaron en estándares preestablecidos, el análisis de los datos ha incluido solamente teoría clásica de los tests y no se ha utilizado un método para el establecimiento del punto de corte.

Por otra parte, el proceso de descentralización llevado adelante por la Udelar desde 2007 que tuvo como corolario la creación de siete centros universitarios en el interior del país, es decir fuera de Montevideo, donde históricamente se concentraron las actividades universitarias, crea oportunidades de acceso a un perfil de estudiantes con ciertas vulnerabilidades, tanto sociales como académicas. Estos estudiantes reúnen características asociadas a la no culminación de los estudios universitarios, tales como dificultades para la finalización educación media, los antecedentes académicos del núcleo familiar y las competencias adquiridas (Rodríguez, 2014). En este sentido, se torna vital conocer las competencias de los estudiantes al ingreso para enfrentar un nivel educativo

[1] Agradecimientos: A los Dres. Ma. Ángeles González Galán y Tabaré Fernández Aguerre directores de la tesis doctoral de la autora "Creación y establecimiento de estándares para la evaluación de la calidad de la educación superior: un modelo adaptado a los Centros Universitarios de la Udelar", defendida en la UNED, donde se produjeron parte de los resultados que se incluyen en este artículo y a los revisores de este artículo por sus valiosos comentarios y sugerencias para mejorarlo.

superior con el fin de adaptar los programas educativos y crear dispositivos de apoyos específicos. Frente a la necesidad de evaluar las competencias de los estudiantes que ingresan a estos centros, se desarrollaron diversas pruebas con este fin (Rodríguez y Correa, 2011; Rodríguez, Correa y Díaz, 2012; Rodríguez, Díaz y Correa, 2013, 2014).

Todas las pruebas referenciadas hasta aquí se basaron en la evaluación de ciertos contenidos, que en algunos casos fueron acordados por los profesores. Sin embargo, se consideró necesario crear pruebas basadas en estándares donde se evaluaran los aprendizajes logrados por los estudiantes, se establecieran metas educativas, y por tanto, que su evaluación fuera coherente con los estándares fijados (Linn y Gronlund, 2000; O'Shea, 2005; Hamilton, Stecher y Yuan, 2008). Ferrer (2006) enfatiza la necesidad de establecer estándares aun cuando existan marcos curriculares nacionales como forma de regular la forma en que se realiza la cobertura de los contenidos y, además, que los criterios de evaluación no resulten dispares entre diferentes contextos.

Por las razones expuestas anteriormente, desde 2014, se comienza a desarrollar pruebas de evaluación basada en estándares, que se aplican por primera vez a la cohorte de estudiantes que ingresaron al siguiente año (Rodríguez et al., 2015). La competencia matemática ha sido considerada como indispensable para los estudiantes que ingresan a la mayoría de las titulaciones y la competencia en lectura se considera transversal a todos los programas educativos. Además, matemática y lectura son consideradas competencias básicas para un estudiante de ingreso a la Universidad (Bertoni, 2005; Zalba et al., 2005). Por estos motivos, se propuso desarrollar pruebas diagnósticas basadas en estándares en estas dos áreas.

El objetivo principal de este artículo es mostrar el proceso de creación y desarrollo de pruebas para la evaluación de competencias básicas (matemática y lectura) al ingreso a la Universidad basadas en estándares, previamente acordados por grupos de expertos, donde se utilizó un método basado en la Teoría de Respuesta al Ítem para el establecimiento de los estándares de desempeño. Además, se presentarán los principales resultados obtenidos a través de estas pruebas y las acciones implementadas para mejorar las carencias detectadas.

2. Fundamentación teórica

El objetivo de las pruebas que nos propusimos elaborar es determinar la posición de los estudiantes en relación con el constructo definido y no su clasificación en función de los individuos que conforman la cohorte. Por este motivo, las pruebas desarrolladas son referidas al criterio y no a la norma. Estas últimas no proporcionan la información necesaria para el diagnóstico de habilidades cognitivas, mientras que el enfoque utilizado se fundamenta en que, a través de los métodos seleccionados, se determine con precisión las competencias de los estudiantes de ingreso. Por eso, el modelo de evaluación elaborado tiene en cuenta una determinada estructura cognitiva, asociada a sus marcos teóricos respectivos –para Matemática el desarrollado por The College Board (2014) y para Lectura el desarrollado por ANEP (2011)- y un conjunto de criterios psicométricos para la elaboración de las pruebas.

La evaluación de tipo diagnóstico en el ámbito educativo tiene su origen en la necesidad de caracterizar a los estudiantes y obtener información que puede ser interpretada por los tomadores de decisiones (Rupp, Templin y Henson, 2010). El principal objetivo es

medir el atributo que se pretende evaluar, por eso se busca desarrollar evaluaciones que sean capaces de determinar con precisión las habilidades cognitivas a evaluar (Gitomer, Steinber y Mislevy, 2009). En nuestro caso, el objetivo de las pruebas diagnósticas desarrolladas es determinar en qué grado los estudiantes superan el constructo definido como competencia en el área de matemática y lectura.

La creación de estas pruebas está sostenida en tres conceptos fundamentales: las pruebas criteriales, los estándares y la interpretación de las puntuaciones basada en la Teoría de Respuesta al Ítem (TRI).

2.1. Pruebas criteriales

Las pruebas referidas al criterio son construidas para medir un determinado dominio de aprendizaje y de esta forma situar a los individuos en relación con ese dominio (Pérez Juste, 2006). Se las diferencia de las pruebas normativas porque mientras estas tratan de ubicar la posición relativa del sujeto con respecto a los demás, las pruebas criteriales tratan de ver en qué medida domina el criterio de referencia (Muñiz, 1998) y permiten conocer con qué grado un estudiante alcanza los niveles de aprendizaje prestablecidos.

El origen de las pruebas criteriales se encuentra en el trabajo de Glaser (1963) donde las define como test elaborados para establecer el nivel de ejecución de un examinado con respecto a un dominio bien definido. Por eso, para la construcción de una prueba criterial debe estar bien definido el universo de medida. El dominio constituye el conjunto de indicadores apropiados para representar el nivel de los sujetos en el constructo que se concretan en ítems (Prieto y Delgado, 1996). Para la creación o adaptación de los ítems se recomienda realizar una revisión desde una perspectiva lógica –a través de jueces que analicen la coherencia entre el ítem y el objetivo a medir– o empírica –la calidad técnica del ítem– (Pérez Juste, 2006).

El desarrollo de pruebas criteriales se fundamentan en dos temas: el análisis y la especificación del dominio y el desarrollo de estándares o un sistema de interpretación de las puntuaciones (Jornet y González Such, 2009).

Según Gil Pascual (2016) las pruebas criteriales se clasifican en función del dominio instruccional de referencia en: test de certificación o admisión (tienen como objetivo decidir sobre la consecución del dominio instruccional de un nivel educativo), test de nivel o dominio (se utilizan para decidir sobre el nivel de instrucción para promover de un curso a otro), test de aula o clase (se circunscribe al aula), test de diagnóstico (pretenden detectar la existencia de problemas relacionados con el aprendizaje), test individualizado (recoge información de aspectos procesuales o de rendimiento de un sujeto).

Las pruebas diagnósticas que se desarrollaron tienen como objetivo determinar con la mayor exactitud posible cuáles son las competencias cognitivas de los estudiantes, por eso se utilizan métodos que pueden aportar esa precisión. Los resultados de estas pruebas se utilizaron para adaptar la enseñanza y el aprendizaje a las necesidades de los estudiantes (Black y Wiliam, 1998) y proveer información para los tomadores de decisiones con el objetivo de planificar políticas educativas.

2.2. Estándares

En las últimas décadas del siglo XX se generaliza la utilización de estándares asociados a la evaluación de aprendizajes (Glass, 1978), especialmente su uso en las pruebas

criteriales (Burton, 1978). En América Latina se comienza a estudiar la relación entre los currículos y las pruebas criteriales a fines de los 90, cuando se conforman grupos de trabajo sobre la temática que recomiendan que los gobiernos establezcan estándares educativos y desarrollen pruebas para medir los resultados (Ferrer, Valverde y Esquivel, 2006).

El concepto de estándares, entendidos en un sentido amplio, se define como un criterio fijo respecto del cual se juzga el resultado o también puede ser comprendido como el logro obtenido (Sotomayor y Gysling, 2011).

En el marco de este trabajo, entendemos los estándares como descriptores de logro, diferenciados por niveles, técnicamente definidos y previamente acordados. Un descriptor debe operacionalizarse para ser evaluado, puede hacerse en términos dicotómicos (presencia/ausencia o disponibilidad/carencia), ordinales (gradación, por ejemplo de acuerdos) o en puntajes (variables métricas).

Un aspecto importante del establecimiento de estándares es alcanzar acuerdos entre los implicados para que sean validados por la comunidad en la que van a ser aplicados. Cizek y Bunch (2007) diferencian entre estándares de contenido y estándares de desempeño. Mientras los primeros son un conjunto de resultados, objetivos curriculares o metas específicas de instrucción que forman el dominio desde el cual se construye un test, el segundo concepto es usado como sinónimo de punto de corte, nivel de logro.

Los estándares de contenido son definidos como la descripción de conocimientos o habilidades específicas sobre la que se espera que los examinados demuestren su dominio acorde a su edad, nivel o campo de estudio (Cizek, Bunch y Koons, 2004) o lo que los estudiantes deben saber y saber hacer en determinadas áreas (Tourón, 2009), es decir los logros de aprendizaje, para demostrar o acreditar ciertos niveles de aprendizajes o competencia. También estos descriptores son tomados como parte de la validez de contenido del instrumento (Hambleton, 2001). Esta diferenciación entre estándares de contenido y estándares de desempeño o rendimiento ha sido la más ampliamente difundida y adoptada (Ravitch, 1996; Tourón, 2009).

Los estándares de desempeño se definen como la descripción del grado de desempeño de los examinados en diferentes categorías (Cizek, Bunch y Koons, 2004) y son usados la mayoría de las veces para informar sobre el desempeño de grupos de estudiantes y el progreso de los centros educativos o estados, en vez de ser utilizados para tomar decisiones acerca de los estudiantes individuales (Linn, 2003).

El objetivo del establecimiento de estándares de desempeño es la clasificación de los estudiantes en niveles. La descripción de los niveles de desempeño provee información sobre los niveles establecidos. Esta descripción es un listado de conocimientos, habilidades o atributos que se consideran integran el nivel de desempeño y que pueden variar en su especificidad (Cizek y Bunch, 2007). Pueden ser planteados con anterioridad al establecimiento de estándares de desempeño (punto de corte) o pueden ser elaborados por el mismo grupo que establezca los estándares de desempeño (Lewis y Green, 1997; Mills y Jaeger, 1988). Ejemplos de descripciones de estándares de desempeño pueden consultarse en Cizek y Bunch (2007), Jornet y González Such (2009) o Tourón (2009).

2.3. Teoría de respuesta al ítem

Con el fin de optimizar la medición de constructos psicológicos o cognitivos y por lo tanto, mejorar la toma de decisiones y resolver ciertos problemas de medición, se ha pasado de la utilización de la Teoría Clásica de los Tests (TCT) a la Teoría de Respuesta al Ítem (TRI). La TRI toma los ítems como unidad de análisis, permite describir algunas propiedades psicométricas del instrumento mediante indicadores invariantes, es decir, que no dependen de la muestra en que se aplique. Esa se puede considerar como su mayor contribución (Muñiz, 1997).

La TRI propone soluciones para las limitaciones de la TCT como la invarianza de los parámetros que permite que el valor de los parámetros de los ítems no depende de la muestra de donde se obtiene; la precisión con la que cada persona es medida según su nivel de rasgo y en función de los ítems concretos que se le hayan aplicado; y los indicadores de bondad de ajuste que permiten estudiar el grado en que los datos ajustan al modelo (Abad et al., 2011).

La TRI se diferencia de la teoría clásica de los tests por utilizar modelos basados en las características de los ítems en vez de las del test, donde las características de los ítems son independientes del grupo en el ítem se ha calibrado y las puntuaciones del rasgo no dependen de las puntuaciones obtenidas en cada test particular. En la TRI se puede obtener una medida de la precisión para cada puntuación del rasgo, lo que la distingue claramente de la teoría clásica y para evaluar la fiabilidad no se requieren de tests estrictamente paralelos.

La TRI establece una relación funcional entre la respuesta del examinado a cada ítem y el rasgo latente responsable de tal realización, es decir, la habilidad, que se nota θ. En la mayoría de los modelos se asume que esta función depende solo de un rasgo, esto es, son unidimensionales. La función que da la probabilidad de obtener determinada puntuación en el ítem condicionado al rasgo, se denomina curva característica del ítem. Las hipótesis que sustentan la TRI son: la unidimensionalidad del espacio latente, la independencia local y la ausencia de factores de velocidad.

3. Métodos

El proceso de trabajo que dio lugar al establecimiento de estándares y creación de instrumentos para evaluar las competencias en Matemática y Lectura se desarrolló durante tres años y se muestra en las figuras 1 y 2.

Se partió de la revisión de los programas curriculares de Matemática de los últimos dos años de bachillerato, estableciéndose como referencia de competencia mínima los contenidos de los programas de las orientaciones de bachillerato con menos contenidos de Matemática. Un grupo de docentes elaboró estándares de contenido agrupados en cuatro categorías: números y operaciones, álgebra y funciones, geometría y análisis de datos, estadística y probabilidad. Esta selección de estándares se presentó a un grupo de discusión integrado por docentes de Matemática de nivel universitario y de Educación media, para ser evaluada. Se realizaron varias rondas de consultas hasta que se acordaron los estándares de contenido sobre los que se basaría la prueba.

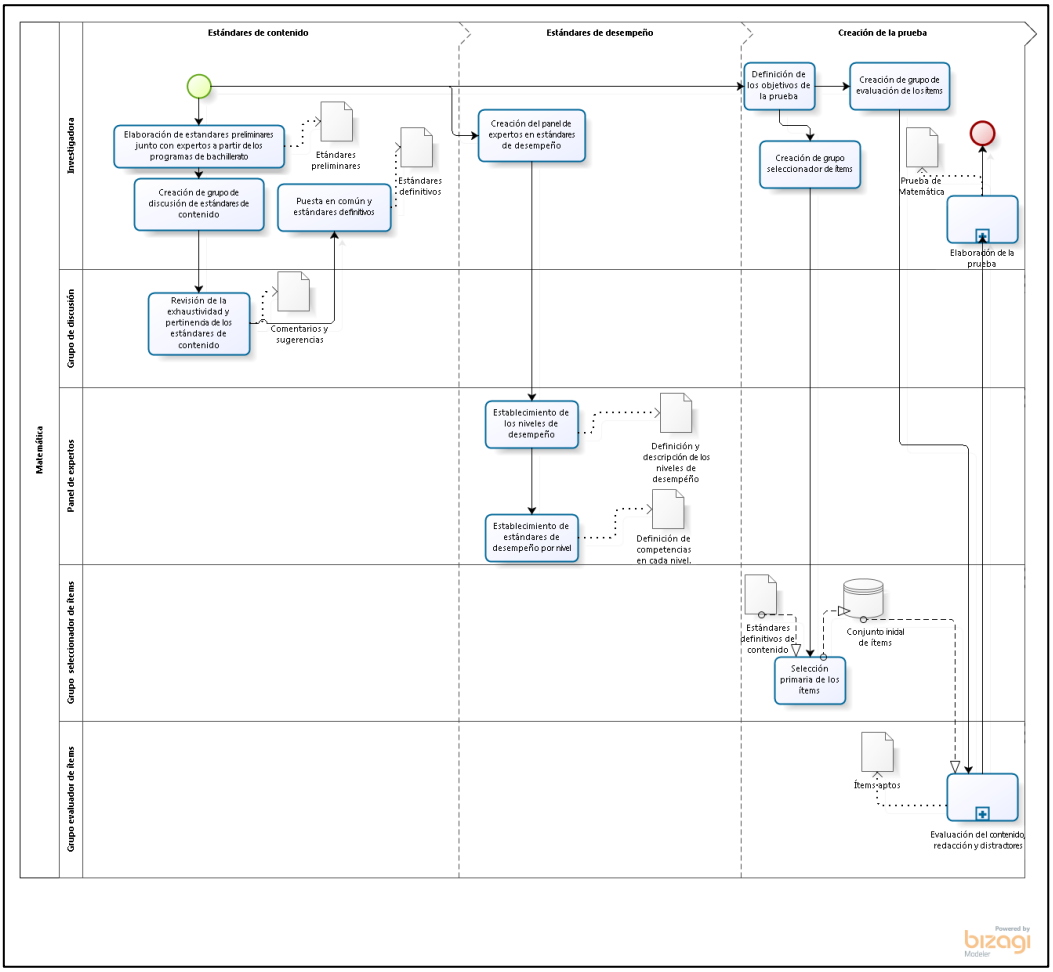


Figura 1. Proceso de establecimiento de estándares y creación de las pruebas de matemática

Fuente: Elaboración propia.

Un panel de expertos, conformado por algunos de los participantes del grupo de discusión que estableció los estándares de contenido más otros especialistas en evaluación de aprendizajes en Matemática, acordó establecer tres niveles de desempeño, que se denominaron Insuficiente, Suficiente y Avanzado y establecieron las competencias que debe tener un estudiante para alcanzar cada uno de los niveles (Rodríguez, 2016). Con la finalidad de contar con ítems ya testeados, se eligió tomar como base la dimensión Matemática del *Scholastic Assessment Test* (SAT) elaborada por The College Board (2014). Un grupo de docentes preseleccionó 186 ítems, repartidos equilibradamente entre las diferentes categorías. Con el objetivo de evaluar la adecuación de los ítems a los estándares de contenido se conformó un grupo de discusión integrado por docentes con experiencia en Educación Media Superior y Universidad y en la evaluación de ítems para pruebas de evaluación en Matemática y se les enviaron los ítems a valorar.

Se consideraron como ítems aptos los que obtuvieron el rango de puntajes más altos dados por los evaluadores. Si había una gran dispersión entre las valoraciones se apeló a los comentarios para decidir si el ítem se consideraba apto.

También se estableció la proporción de ítems para cada categoría teniendo en cuenta la preponderancia que se les da en los programas curriculares a esos contenidos. Se determinó que cada categoría podía tener una proporción de ítems distribuidos de la siguiente forma.

Tabla 1: Proporción de ítems de Matemática en cada categoría

CATEGORÍA	PROPORCIÓN DE ÍTEMS
Números y Operaciones	15-18%
Álgebra	45-52%
Geometría	15-20%
Análisis de datos, estadística y probabilidad	9-14%

Fuente: Rodríguez (2016).

Finalmente, se eligieron 88 ítems para la prueba de matemática que conformaron dos cuadernillos con 44 ítems cada uno. La proporción final de ítems por categoría así como la tabla de especificaciones se puede leer en Rodríguez (2016). En la fig. 1 se muestra el proceso completo de establecimiento de estándares y creación de la prueba de matemática.

Los estándares de contenido para la prueba de lectura se elaboraron siguiendo las "Pautas de referencia sobre los niveles de lectura en español como primera lengua" del Programa de Lectura y Escritura en Español (ANEP, 2011). Estas pautas establecen categorías parametrizadas que permiten describir los conocimientos y aptitudes lectoras de los estudiantes. La categoría L4B que corresponde con el lector que finalizó el bachillerato, en sus tres categorías: componentes de lectura, conocimiento lingüístico y géneros discursivos es la que se utiliza como referencia. En la figura 2 se puede apreciar el proceso que dio lugar a la prueba de lectura.

También para la prueba de lectura se utilizaron ítems testeados previamente de tres fuentes diferentes: prueba de evaluación diagnóstica en la dimensión Lectura de Rodríguez, Díaz y Correa (2014), ítems liberados de PISA 2009 e ítems elaborados por el Programa de Lectura y Escritura Académica (LEA) de la Comisión Sectorial de Enseñanza de la Universidad de la República y aplicados a la generación de ingreso 2014.

Se estableció que la proporción adecuada de ítems por cada categoría fuera distribuida de la siguiente forma.

Tabla 2. Proporción de ítems de Lectura en cada categoría.

CATEGORÍA	PROPORCIÓN DE ÍTEMS
Componentes de Lectura	60-75%
Conocimiento linguístico	15-20%
Géneros discursivos	1-5%

Fuente: Rodríguez (2016).

Para la selección de los ítems se utilizó la información disponible acerca de sus propiedades y se procedió a adaptar los ítems que lo requirieran. La prueba de lectura quedó conformada por 37 ítems.

En la siguiente figura 2 se muestra el proceso de creación de la prueba de Lectura.

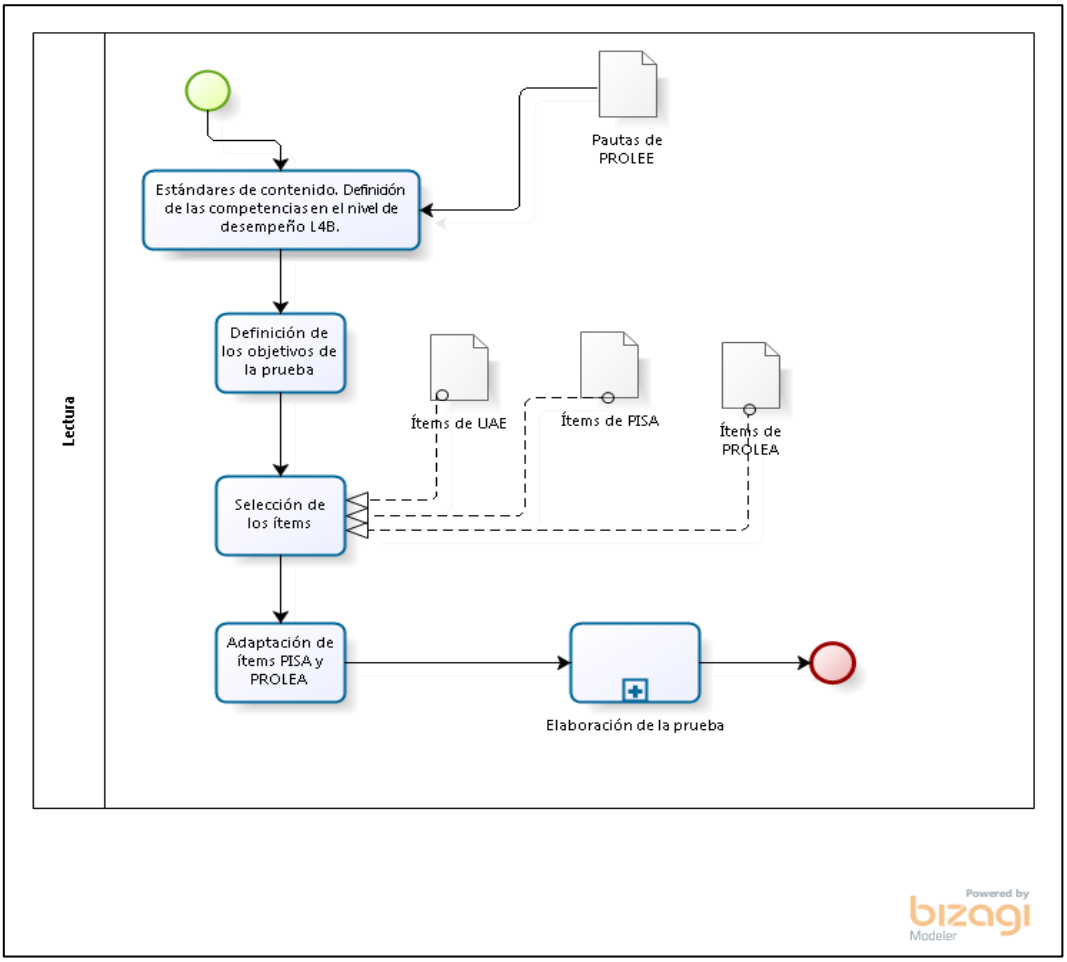


Figura 2. Proceso de creación de la prueba de lectura
Fuente: Elaboración propia.

3.1. Muestra

Se aplicó la prueba de Matemáticas, con sus dos cuadernillos y la prueba de Lectura a los estudiantes que ingresaron a los Centros Universitarios Regionales (CENUR) de la Udelar (Centros Universitarios Regional del Este, Litoral Norte y Noreste, que incluyen las sedes de Maldonado, Paysandú, Rivera, Rocha, Salto, Tacuarembó, Treinta y Tres y Melo) en 2015. La tasa global de respuesta fue del 65,8%. Aunque se apuntaba a un censo, los dos meses que transcurrieron entre la inscripción a la Universidad y la aplicación de la primera prueba hizo que una proporción de los estudiantes hubiera decidido no comenzar los cursos, y por lo tanto, vieron innecesario cumplir con la realización de las pruebas. La evaluación de matemática, realizada en primer lugar, fue cumplimentada por 1.380 estudiantes y la de lectura, aplicada en segunda instancia, fue realizada por 1244 estudiantes. Como se usa TRI no es necesario que la muestra sea representativa por el principio de invarianza.

Un análisis detallado de la cobertura y de las características de la muestra se pueden leer en Rodríguez, Figueroa y Fernández (2016).

Las pruebas se aplicaron a través de la plataforma del Entorno Virtual de Aprendizaje (EVA) de la Udelar. Este espacio proporciona la posibilidad de realizar una prueba fija informatizada.

4. Resultados

Se realizaron una serie de análisis psicométricos a efectos de estudiar las propiedades de las pruebas. Se consideraron para este estudio aquellos estudiantes que no tuviesen más de cinco ítems omitidos.

En primer lugar, se calcularon los índices de dificultad e índices de discriminación de cada ítem así como la distribución de las respuestas.

La dificultad media de los ítems del cuadernillo 1 de matemática fue de 0,36 (0,02), del cuadernillo 2 fue 0,37 (0,02) y en la prueba de lectura se encontró en 0,52 (0,03). Se recomienda que el índice de dificultad sea mayor a 0,5 y que la dificultad media sea 0,62 para ítems de 4 opciones (Abad et al., 2011). Si bien los ítems de nuestras pruebas tienen 5 distractores, nos encontramos alejados de lo recomendado para el índice de dificultad en las pruebas de Matemática y en el rango en la de Lectura. Hay que resaltar que las pruebas de Matemática fueron elaboradas siguiendo estándares que contemplan los contenidos mínimos en Matemática para los egresados de los bachilleratos. Esto nos demuestra que, aunque los ítems evalúan competencias básicas, resultaron difíciles a la muestra en que se aplicó la prueba.

Fueron calculados los índices de discriminación, en este caso las correlaciones biseriales puntuales (bivariadas) de los ítems por separado y de los grupos. La discriminación media de los ítems del cuadernillo 1 y 2 de Matemática fue de 0,40 (0,01 y 0,02 respectivamente) y para la prueba de Lectura se situó en 0,33 (0,01). La bibliografía sugiere que los ítems que obtengan un índice de discriminación menor a 0,20 sean descartados (Abad et al., 2011; Schmeiser y Welch, 2006). Este criterio se usó para la eliminación de los ítems que no ajustaban. En esta primera instancia no se eliminaron ítems del cuadernillo 1 de matemática; se sacaron los ítems 9, 11, 19, 28 y 31 del cuadernillo 2 de Matemática y se eliminaron los ítems 9, 19, 29 y 34 de la prueba de Lectura. Una vez eliminados estos ítems se calculó el coeficiente alfa de Cronbach para determinar la fiabilidad de las pruebas.

El alfa de Cronbach de las pruebas de Matemática fueron 0,88 para el cuadernillo 1 y 0,87 para el cuadernillo 2. El alfa de la prueba de Lectura se situó en 0,75. (Rodríguez et al., 2015). Los coeficientes de fiabilidad son muy buenos para la prueba de Matemática y bueno para la prueba de Lectura.

Para estudiar la validez se analizó la unidimensionalidad de las pruebas. Con la finalidad de conocer si la matriz de datos es adecuada para realizar un análisis factorial se halla el índice Kaiser-Meyer-Olkin y se efectúa el test de esfericidad de Barlett.

El valor global del índice de adecuación muestral de KMO para el cuadernillo 1 de Matemática es de 0,913 y de 0,899 para el cuadernillo 2, que según la clasificación de Kaiser se puede considerar como "maravillosa". Para la prueba de Lectura el índice de KMO es de 0,836, que se considera "meritoria".

En el test de esfericidad de Barlett el valor del chi-cuadrado para el cuadernillo 1 de Matemática es 5524,307(946) p-value=0, para el cuadernillo 2 de Matemática es

4643,2(946) p-value=0 y para el cuadernillo de Lectura es 3697,972(666) p-value=0. Por lo tanto, en los tres casos podemos rechazar la hipótesis nula de que las variables utilizadas en el análisis no se correlacionarían en la población en la que se ha extraído la muestra.

A partir de los dos estudios anteriores podemos concluir que los datos son adecuados para realizar un análisis factorial. Se realiza un análisis de componentes principales con rotación varimax para determinar si las dimensiones a priori son unidimensionales.

Con el propósito de evaluar la unidimensionalidad se utilizó el criterio propuesto por Reckase (1979) donde se pide que el primer componente explique, por lo menos, el 20 por ciento de la varianza. El porcentaje de varianza explicada por el primer factor es 26,37% y el ratio entre los dos primeros valores propios es 3,85 para el cuadernillo 1 de Matemática. Para el cuadernillo 2 de Matemática la varianza explicada por el primer factor es 26,69% y el ratio entre los dos primeros valores propios es 4,92. En la prueba de Lectura la varianza explicada por el primer factor es 18,46% y el ratio entre los dos primeros valores propios es 3,98. Si bien en este último caso la varianza del primer factor no alcanza al 20%, se destaca que el ratio entre los dos primeros valores propios es alto.

También se utilizó el cociente entre el primer autovalor y el segundo. Este cociente será indicativo de unidimensionalidad si es aproximadamente 4. En el cuadernillo 1 de Matemática el cociente entre el primer autovalor y el segundo es de 3,84, en el cuadernillo 2 de Matemática es de 4,91 y en el cuadernillo de Lectura es de 3,98.

A partir de estos análisis realizados se puede considerar que las pruebas son unidimensionales y, por lo tanto, es posible aplicar la TRI.

En primer lugar, se calculan las curvas características de los ítems (CCI) y los parámetros de dificultad y discriminación de cada cuadernillo de Matemática y de la prueba de Lectura.

Para calibrar nuestra prueba nos basamos en la TRI y se utilizó el modelo logístico de 2 parámetros. Se estudió el ajuste al modelo tanto de las curvas características del ítem (CCI) como de la habilidad de las personas (*person fit*). Para medir el ajuste de las personas usamos el estadístico L_o de Levine y Rubin (1979) y su versión estandarizada L_z propuesta por Drasgow et al. (1985). A través de este análisis se encontró que el 96,45 % de los sujetos ajustaron bien en el cuadernillo 1 de Matemática, el 98,46 % ajustaron bien para el cuadernillo 2 y el 96,44% para el cuadernillo de Lectura (Rodríguez, 2016).

Se equipararon los cuadernillos 1 y 2 de Matemática mediante ítems de anclaje usando el método de media y desviación (Hambleton, Swaminathan y Rogers, 1991). Los coeficientes para llevar los ítems de la escala del cuadernillo 1 a la escala común fueron α =0,91 y β=0,13. Para el cuadernillo 2 fueron α =1,08 y β =-0,05.

4.1. Establecimiento del punto de corte

La elección de un método para el establecimiento del punto de corte es un tema crucial en el establecimiento de estándares y en la interpretación de los resultados (Jornet, González Such y Suárez, 2010). Con el objetivo de lograr la mayor independencia entre la dificultad empírica y los puntos de corte se optó por el método propuesto por García

et al. (2013), ya que los ítems son diseñados, o en nuestro caso, seleccionados en base a los estándares de desempeño establecidos para clasificar a los estudiantes.

Este nuevo método, basado en la TRI, consta de cinco pasos que se sintetizan a continuación:

1. Construcción o selección de un banco de ítems basado en los estándares de contenido.
2. Calibración del banco de ítems y estimación de las curvas características de los ítems (CCI). En nuestro caso se utilizó el modelo logístico de 2 parámetros.
3. Cálculo de las CCI promedio para cada familia de ítems (para todos los ítems que se encuentran en el mismo nivel de desempeño).
4. Cálculo de las CCI promedio conjuntas para cada familia de ítems.
5. Cálculo del punto de corte. Para nuestras pruebas se calcularon 3 puntos de corte de forma de clasificar a los sujetos en los diferentes niveles de desempeño.

Los estudiantes quedaron clasificados según el θ obtenido de la siguiente forma.

Tabla 3. Clasificación de los estudiantes según la habilidad obtenida en las pruebas de Matemática

θ	NIVEL
<0.62	Insuficiente
>=0.62 y <2.06	Suficiente
>=2.06	Avanzado

Fuente: Rodríguez (2016).

4.2. Desempeño en Matemática

El 78,06% de los estudiantes quedaron en el nivel Insuficiente, el 18,3% en el nivel Suficiente y el 3,65% en el nivel Avanzado. Estos resultados son coherentes con los obtenidos a través de otra prueba de evaluación diagnóstica en Matemática para estudiantes de ingreso a la Universidad, esto es, distintos ítems y diferente método para el cálculo del punto de corte. En ese caso se utilizó una aplicación del método de Angoff para clasificar a los estudiantes en dos niveles (suficiente e insuficiente). Se compararon los resultados con la dimensión Resolución de problemas, que evaluó competencias matemáticas para el nivel de egresados de bachillerato (Rodríguez, Díaz y Correa, 2014). La proporción de suficientes estuvo en el 22%. Si sumamos el nivel de desempeño suficiente y avanzado de los resultados obtenidos en 2015 con los instrumentos creados, obtenemos un 21,95% de estudiantes que superaron la suficiencia. También los resultados hallados en otras pruebas de matemática para estudiantes de ingreso a la Universidad presentan proporciones de suficiencia igualmente bajos (Míguez, Blasina y Loureiro, 2013; Mussio y Martinotti, 2013; UEFI, 2012).

Las competencias asociadas al nivel de desempeño Insuficiente indican que el 78,06% son capaces de resolver problemas de aritmética incluyendo porcentajes, razones y proporciones, pueden operar con fracciones, sustituyen y simplifican expresiones algebraicas simples, pueden determinar el límite de una función dada por su gráfica. La mayoría de estas competencias se adquieren durante el primer ciclo de la Educación Media. Algunas de las competencias que se incluyen dentro del nivel Suficiente, alcanzado por el 18,3% de los estudiantes, muestran que son capaces de operar con fracciones usando paréntesis, operar con números complejos y representarlos en el plano

complejo, resolver ecuaciones y desigualdades complejas de una variable real, resolver problemas de conteo utilizando números combinatorios, operar con exponenciales, logaritmos y potencias, calcular la media, mediana, moda, cuartiles, varianza o resolver problemas usando las propiedades de la probabilidad. Estas competencias se corresponden con las que se deben adquirir al término del segundo ciclo de la Educación Media en las orientaciones de bachillerato con menos contenidos de Matemática. La definición completa de las competencias en matemática en cada nivel de desempeño se encuentra en Rodríguez (2016).

4.3. Desempeño en Lectura

Se obtuvieron los θ para clasificar a los estudiantes en los distintos niveles de desempeño. Se presenta en la tabla 4.

Tabla 4. Clasificación de los estudiantes según la habilidad obtenida en la prueba de Lectura

θ	NIVEL
<-0.085	Insuficiente
>=-0.085 y <1.85	Suficiente
>=1.85	Avanzado

Fuente: Rodríguez (2016).

El 46,98% de los estudiantes quedó clasificado en el nivel Insuficiente, 51,18% en el Suficiente y el 1,84% en el Avanzado. Si comparamos estos resultados en Lectura con los obtenidos en la dimensión Comprensión y aplicación de la prueba aplicada en 2012, también podemos afirmar que existe coherencia, ya que en esa instancia hubo un 40% de suficientes (Rodríguez, Díaz y Correa, 2014), mientras que en esta prueba un 53% superaron la suficiencia.

Las competencias de los estudiantes que quedaron clasificados en el nivel Insuficiente incluye relacionar al autor y al texto con su contexto histórico y sociocultural, establecer relaciones intratextuales entre los diferentes bloques del texto, reconocer las estructuras sintácticas del español. Estas competencias están asociadas a los estudiantes que culminaron el primer ciclo de la Educación Media. El 51,18% de los estudiantes que alcanzaron el nivel de Suficiente poseen una lectura reflexiva y crítica, reconocen posibles inconsistencias internas de los textos, reconocen el léxico básico de las disciplinas específicas asociadas con la educación formal. Las competencias de este nivel se asocian a estudiantes que culminaron el segundo ciclo de la Educación Media. La descripción completa de cada nivel de lector se encuentra en ANEP (2011).

5. Discusión

En primer lugar, debemos resaltar el desarrollo metodológico de las pruebas donde se siguió lo recomendado por la literatura especializada para el establecimiento de estándares de contenido y desempeño (Cizek y Bunch, 2007; Jornet, González Such y Suárez, 2010; Tourón, 2009).

Se obtuvieron dos pruebas calibradas mediante TRI y validadas para la evaluación de competencias en matemática y lectura para el nivel de ingreso a la universidad. La validez de contenido de los instrumentos está fundamentada en el procedimiento utilizado para la delimitación de los contenidos a evaluar, es decir, el propio

establecimiento de estándares de contenido. La validez de constructo, obtenida a través de los análisis factoriales, se aprecia en la concordancia entre el contenido de los estándares y los factores obtenidos.

Dentro de los aspectos metodológicos se destaca la utilización de un método para el establecimiento de los puntos de cortes basados en la TRI que aporta solidez al análisis y a la interpretación de los resultados (Muñiz, 1997; Jornet, González Such y Suárez, 2010).

La baja proporción de estudiantes que superaron el nivel de suficiencia es congruente con otras evaluaciones y pone evidencia la vulnerabilidad académica de los estudiantes de los CENUR, que captan mayoritariamente estudiantes locales, es decir, que viven en la capital departamental donde se ubica la Sede, o de la región, provenientes de los pueblos cercanos. La comparación del desempeño en Matemática entre estudiantes de ingreso provenientes de Montevideo contra los del Interior realizada por Mussio y Martinotti (2013) apunta en el mismo sentido, ya que los estudiantes que realizaron la educación preuniversitaria en el Interior obtuvieron menores puntajes en la prueba, es más, cuanto más lejos de la capital se encontraba la residencia del estudiante menor era el puntaje obtenido. Los bajos resultados, tanto en Matemática como Lectura, se corresponden con los hallazgos de la investigación sobre "Trayectorias académicas y laborales de los jóvenes uruguayos - El panel PISA 2003-2007" de Boado y Fernández (2010). Los estudiantes con menores puntajes que lograron acceder a la educación superior fueron los que cursaron educación media en el Interior del país.

Por otra parte, la devolución de resultados a los estudiantes, obteniendo una retroalimentación inmediata sobre sus aciertos y errores, además de un comentario global con sugerencias según el nivel de desempeño que se encontraban, les permitió elegir trayectorias alternativas o tomar cursos compensatorios que se crearon especialmente para apoyar las dificultades detectadas. Las investigaciones realizadas en estudiantes con baja preparación en matemática para afrontar las exigencias universitarias indican que los cursos de apoyo o de nivelación mejoran los resultados (Bettinger y Long, 2009; Hillock et al., 2013; Perkin y Bamforth, 2009).

6. Conclusiones

Sobre los resultados en las pruebas se puede concluir que, teniendo en cuenta los estándares de contenido y la complejidad de los ítems, las proporciones de estudiantes que lograron la suficiencia son muy bajos en Matemáticas (22%) y bajos en Lectura (53%). La coherencia de estos resultados con los obtenidos en otras pruebas con objetivos similares proporciona evidencia sobre la solidez de los instrumentos (Mussio y Martinotti, 2013; Rodríguez, Díaz y Correa, 2013, 2014; UEFI, 2012).

Una mención especial merece la definición de políticas educativas en función de los resultados obtenidos, ya que era uno de los objetivos planteados. Se implementaron dos programas compensatorios en el área de Matemática, uno en la región Noreste y otro en el CENUR Este, como forma de apoyo para los estudiantes que fueron clasificados en el nivel Insuficiente. Por otra parte, el Programa de Lectura y Escritura Académica de la Comisión Sectorial de Enseñanza impartió talleres de apoyo en lectura para todos los estudiantes que lo requirieran. Algunas de las primeras políticas implementadas se describen en Rodríguez, Figueroa y Fernández (2016).

Además, desde la Comisión Coordinadora del Interior se han impulsado varias líneas de acción: se institucionalizó la Evaluación Diagnóstica como un programa y su Plenario resolvió continuar la aplicación de las pruebas para 2016; se resolvió integrar un equipo académico en Lectura y Escritura para elaborar un programa de apoyo en Lectura (CCI, 2016a); se impulsó trabajo conjunto con la Comisión Sectorial de Enseñanza (CCI, 2016b) y se conformó un Grupo Académico de Matemática con el doble objetivo de elaborar un programa de apoyo en Matemática y colaborar con la creación y evaluación de ítems para la prueba diagnóstica (CCI, 2016c). Este grupo produjo un programa de apoyo en Matemática con características innovadoras, que es recibido por el Plenario de la CCI, encomendando a los coordinadores y comisiones de carrera del Interior estudiar la propuesta para su implementación (CCI, 2016d).

Referencias

Abad, F., Olea, J., Ponsoda, V., García, C. (2011). *Medición en Ciencias Sociales y de la Salud.* Madrid: Síntesis.

Altmark, S., Castrillejo, A., Debera, L. y Nalbarte, L. (2006). *Elaboración de pruebas diagnósticas al ingreso a la Facultad de Ciencias Económicas y Administración DT (06/02).* Recuperado de http://www.iesta.edu.uy/wp-content/uploads/2010/03/0602.pdf

ANEP. (2011). *Pautas de referencias sobre niveles de lectura en español como primera lengua.* Montevideo: PROLEE, Administración Nacional de Educación Pública.

Attewell, P., Heil, S. y Reisel, L. (2012). What is academic momentum? And does it matter? *Educational Evaluation and Policy Analysis, 34*(1), 27-44. doi:10.3102/0162373711421958

Bertoni, E. (2005). *El estudiante universitario: Una aproximación al perfil de ingreso. Documento de trabajo No. 3.* Montevideo: Unidad Académica, Comisión Sectorial de Enseñanza.

Bettinger, E. P. y Long, B. T. (2009). Addressing the needs of under-prepared students in higher education: Does college remediation work? *The Journal of Human Resources, 44,* 736-771. doi:10.1353/jhr.2009.0033

Black, P. y Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Boado, M. y Fernández, T. (2010). *Trayectorias académicas y laborales de los jóvenes en Uruguay. El panel PISA 2003-2007.* Montevideo: FCS-UDELAR.

Burton, N. W. (1978). Societal standards. *Journal of Educational Measurement, 15*(4), 263-271. doi:10.1111/j.1745-3984.1978.tb00073.x

Cizek, G. J. y Bunch, M. B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests.* Thousand Oak, CA: Sage Publications.

Cizek, G. J., Bunch, M. B. y Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-31. doi:10.1111/j.1745-3992.2004.tb00166.x

CollegeBoard. (2014). *SAT Practice Questions Math.* Recuperado de http://sat.collegeboard.org/practice/

Comisión Coordinadora del Interior. (2016a). *Resolución No. 28 del 10/02/2016.* Recuperado de http://www.expe.edu.uy/expe/resoluci.nsf/ed0853334a4fde3c83257c8d007f3233/426a58292b473fb503257f73004431bd?OpenDocument

Comisión Coordinadora del Interior. (2016b). *Resolución No. 704 del 29/08/2016)*. Recuperado de http://www.expe.edu.uy/expe/resoluci.nsf/ed0853334a4fde3c83257c8d007f3233/5696d1b256ae8a15032580360052db19?OpenDocument

Comisión Coordinadora del Interior. (2016c). *Resoluciones 132 y 133 del 07/03/2016)*. Recuperado de http://www.expe.edu.uy/expe/resoluci.nsf/ed0853334a4fde3c83257c8d007f3233/a586a2373f0d542b03257f9600608060?OpenDocument

Comisión Coordinadora del Interior. (2016d). *Resolución No. 1069 del 12/12/2016*. Recuperado de http://www.expe.edu.uy/expe/resoluci.nsf/ed0853334a4fde3c83257c8d007f3233/b1d92c14f5bf21ea0325808b006734eb?OpenDocument

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E. y Candell, G. L. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement*, *13*(3), 285-299. doi:10.1177/014662168901300309

Enrich, H., Míguez, M., Rodríguez Ayán, M. N. y Leymonié, J. (2006). *Evaluación diagnóstica de las habilidades matemáticas al ingreso en las facultades del área científico-tecnológica*. Montevideo: CSE, Udelar.

Ferrer, G. (2006). *Estándares en educación. Implicancias en América Latina*. Santiago de Chile: PREAL.

Ferrer, J. G., Valverde, G. y Esquivel, J. M. (2006). *Aspectos del currículo prescrito en América Latina: Revisión de tendencias contemporáneas en currículo, indicadores de logro, estándares y otros instrumentos. Informe de trabajo*. Chile: PREAL.

García, P. E., Abad, F. J., Olea, J. y Aguado, D. (2013). A new IRT-based standard setting method: Application to elath-listening. *Psicothema*, *25*(2), 238-244.

Gil Pascual, J. A. (2016). *Técnicas e instrumentos para la recogida de información*. Madrid: UNED.

Gitomer, D. H., Steinber, L. S. y Mislevy, R. J. (2009). Diagnostic assessment and troubleshooting skill in an intelligent tutoring system. En P. D. Nichols, S. F. Chipman y R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 73-102). Mahwah, NJ: Lawrence Erlbaum Associates.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American psychologist*, *18*(8), 519-521. doi:10.1037/h0049294

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*(4), 237-261. doi:10.1111/j.1745-3984.1978.tb00072.x

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Nueva York, NY: Lawrence Erlbaum.

Hambleton, R., Swaminathan, H. y Rogers, H. (1991). *Fundamentals of Item Response Theory*. Thousand Oaks, CA: Sage Publications.

Hamilton, L. S., Stecher, B. M. y Yuan, K. (2008). *Standards-based reform in the United States: History, research and future directions*. Los Ángeles, CA: RAND Corporation.

Hillock, P. W., Jennings, M., Roberts, A. y Scharaschkin, V. (2013). A Mathematics support programme for first-year engineering students. *International Journal of Mathematical Education in Science and Technology*, *44*(7), 1030-1044. doi:10.1080/0020739x.2013.823251

Jornet, J. y González Such, J. (2009). Evaluación criterial: Determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios sobre Educación*, *16*, 103-121.

Jornet, J., González Such, J. y Suárez, J. (2010). Validación de los procesos de determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios sobre Educación, 19*, 11-29.

Levine, M. V. y Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics, 4*(4), 269-290. doi:10.2307/1164595

Lewis, D. M. y Green, D. R. (junio, 1997). The validity of performance level descriptors. Comunicación presentada en el *Annual CCSSO Conference on Large Scale Assessment*, Universidad de Colorado, Springs, CO.

Linn, R. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Recuperado de http://epaa.asu.edu/epaa/v11n31/

Linn, R. L. y Gronlund, N. E. (2000). *Measurement and assessment in teaching*. Upper Saddle River, NJ: Prentice-Hall.

Míguez, M., Blasina, L. y Loureiro, S. (septiembre, 2013). *Diagnóstico al ingreso en la Facultad de Ingeniería de la Universidad de la República: Matemática y variables no tradicionales.* Comunicación presentada en el VII Congreso Iberoamericano de Educación Matemática, Montevideo. Recuperado de http://www.cibem7.semur.edu.uy/7/actas/pdfs/873.pdf

Mills, C. N. y Jaeger, R. M. (1988). Creating descriptions of desired student achievement when setting performance standards. En L. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 73-86). Washington D. C.: Council of Chief State School Officers.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide.

Muñiz, J. (1998). *Teoría Clásica de los Tests*. Madrid: Pirámide.

Mussio, I. y Martinotti, L. (2013). *Informe sobre prueba diagnóstica aplicada a estudiantes que ingresan a la Facultad de Ciencias Sociales.* Recuperado de http://cienciassociales.edu.uy/departamentodeeconomia/wp-content/uploads/sites/2/2013/archivos/1213.pdf

O'Shea, M. (2005). *From standards to success*. Alexandria, VA: Association for Supervision and Curriculum Development.

Pérez Juste, R. (2006). *Evaluación de programas educativos*. Madrid: La Muralla.

Perkin, G. y Bamforth, S. (2011). A variety of approaches to the provision of mathematics help for first-year engineering undergraduates. *International Journal of Electrical Engineering Education, 48*(1), 80-91. doi:10.7227/ijeee.48.1.7

Prieto, G. y Delgado, A. (1996). Construcción de ítems. En J. Muñiz (Coord.), *Psicometría* (pp. 105-138). Madrid: Universitas.

Ravitch, D. (1996). *Estándares Nacionales en Educación*. Santiago de Chile: PREAL.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics, 4*(3), 207-230. doi:10.2307/1164671

Rodríguez, P. (2014). Oportunidades y riesgos en el acceso a la Educación Superior en el marco del Centro Universitario de la Región Este. En T. Fernández y A. Ríos (Eds.), *El tránsito entre ciclos en el Educación Media y Superior de Uruguay* (pp. 165-181). Montevideo: CSIC.

Rodríguez, P. (2016). *Creación y establecimiento de estándares para la evaluación de la calidad de la educación superior: Un modelo adaptado a los centros universitarios de la Udelar* (Tesis doctoral). Universidad Nacional de Educación a Distancia, Madrid.

Rodríguez, P. y Correa, A. (2011). *Informe de las Evaluaciones Estudiantiles 2011. Evaluaciones de centro, docente y asignaturas.* Maldonado: CURE.

Rodríguez, P., Correa, A. y Díaz, M. (2012). *Informe sobre Evaluación Diagnóstica 2012.* Maldonado: CURE.

Rodríguez, P., Díaz, M. y Correa, A. (2013). *Resultados de la Evaluación Diagnóstica 2013.* Maldonado: CURE.

Rodríguez, P., Díaz, M. y Correa, A. (2014). Los aprendizajes al ingreso en un Centro Universitario Regional. *Intercambios, 2*(1), 91-100.

Rodríguez, P., Carreño, G., Fernández, T., Figueroa, V. y Lorda, N. (2015). *Evaluación diagnóstica 2015 en Matemática y Lectura. Primer informe de resultados.* Montevideo: CCI.

Rodríguez, P., Figueroa, V. y Fernández, T. (2016). Evaluación de competencias al ingreso a la Universidad. *Temas de Educación, 22*(1).

Rupp, A., Templin, J. y Henson, R. (2010). *Diagnostic measurement. Theory, methods and applications.* Nueva York, NY: The Guilford Press.

Schmeiser, C. B. y Welch, C. J. (2006). Test development. En R. L. Brennan (Ed.), *Educational measurement* (pp. 307-353). Westport, CT: Praeger Publishers.

Sotomayor, C. y Gysling, J. (2011). Estándares y regulación de calidad de la formación de profesores: Discusión del caso chileno desde una perspectiva comparada. *Calidad en la Educación.* (35), 91-129. doi:10.4067/s0718-45652011000200004

Tourón, J. (2009). El establecimiento de estándares de rendimiento en los sistemas educativos. *Estudios sobre Educación, 16*, 127-146.

Unidad de Enseñanza de Facultad de Ciencias. (2005). *Evaluación Diagnóstica de Conocimientos y Habilidades al Ingreso.* Recuperado de http://ue.fcien.edu.uy/Documentos/Edich2005.pdf

Unidad de Enseñanza de Facultad de Ciencias. (2010). *Evaluación Diagnóstica de Conocimientos y Habilidades al Ingreso.* Recuperado de http://ue.fcien.edu.uy/Documentos/EDICH_2010_primer_informe.pdf

Unidad de Enseñanza de Facultad de Ingeniería. (UEFI). (2012). *Informe de la Herramienta Diagnóstica al Ingreso.* Recuperado de https://www.fing.edu.uy/sites/default/files/noticias/2012/7/5793/Informe%20final%20%20HDI2012.pdf

Zalba, M. E., Gómez de Erice, M. V., Alfonso, V., Deamici, C., Erice, X., Gutiérrez, N. B., … Sayavedra, C. (2005). *Competencias para el ingreso y permanencia en la universidad: Una propuesta de articulación curricular entre el nivel superior y el nivel medio de enseñanza: La experiencia de la Universidad Nacional de Cuyo.* Recuperado de http://www.cinda.cl/download/libros/Curr%C3%ADculo%20Universitario%20Basado%20en%20Competencias.pdf

Breve CV de la autora

Pilar Rodríguez Morales

Doctora en Educación por la UNED, España. Profesora Adjunta, Coordinadora de la Unidad de Apoyo a la Enseñanza del Centro Universitario Regional del Este, Universidad de la República de Uruguay. Corresponsable del Programa de Evaluación

Diagnóstica de la Comisión Coordinadora del Interior de la Universidad de la República. Integrante del Sistema Nacional de Investigadores desde 2016. ORCID ID: 0000-0003-1929-4961. Email: prodriguez@cure.edu.uy

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 109-129.
https://doi.org/10.15366/riee2017.10.1.006

Desarrollo y Validación de un Instrumento para Evaluar la Práctica Docente en Educación Preescolar

Development and Validation of an Instrument for Evaluating the Teaching Practice in Early Childhood Education

Luis Horacio Pedroza Zúñiga *
Edna Luna

Universidad Autónoma de Baja California

Actualmente se reconoce que la calidad de la enseñanza contribuye a elevar los aprendizajes de los niños, por ello se tiene un interés creciente en evaluar de forma eficiente y precisa lo que realizan los docentes en las aulas. Este artículo da cuenta de las evidencias de validez de contenido y constructo de un instrumento para evaluar la práctica docente en educación preescolar, así como la consistencia de la medición y del proceso de observación. El instrumento en su versión más acabada se compone de 30 rúbricas divididas en tres dimensiones: planeación, intervención y evaluación. Para ello se analiza la planeación didáctica y la práctica docente en el aula. La información brindada por el instrumento logró una alta consistencia interna y buen ajuste al modelo de Rasch. El análisis factorial exploratorio permitió identificar tres factores subyacentes, con diferencias respecto a la estructura teórica. Los análisis para identificar la confiabilidad entre los observadores, Kappa cuadrática ponderada y análisis de la varianza a partir de la Teoría de la Genereralizabilidad ofrecieron resultados coincidentes: adecuada consistencia en las distintas observaciones.

Palabras clave: Evaluación docente, Educación preescolar, Observación sistemática, Rúbricas, Confiabilidad, Validez.

This article reports the content and construct-related validity evidence of an instrument for evaluating teaching practice in early childhood education. In addition, it shows evidence of the consistency of measurement and observation process. The instrument consists of 30 rubrics divided into three dimensions: planning, intervention and evaluation. To do this, the planning and the teaching practice is analyzed. The information provided by the instrument achieved a high internal consistency and a good fit to the Rasch model. Exploratory factor analysis identified three underlying factors, with differences in the theoretical structure. Analyses to identify reliability among observers, quadratic weighted Kappa and analysis of variance based on the generalizability theory, showed consistent results: proper consistency among the different observations. Further analyses are needed with a larger sample of examinees, in order to confirm the factorial structure and adjust the instrument to increase its accuracy.

Keywords: Teacher evaluation, Early childhood education, Measurement of instruction, Rubrics, Reliability, Validity.

*Contacto: horaciopedroza@hotmail.com

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 4 de julio de 2016
1ª Evaluación: 9 de agosto de 2016
Aceptado: 1 de septiembre de 2016

1. Introducción

Evaluar la enseñanza que realizan los docentes en las aulas es una tarea demandante. Esto es así porque la enseñanza es una actividad multidimensional y compleja. La profesión docente requiere del dominio de competencias especializadas, es necesario un conocimiento de lo que se enseña, de cómo enseñar y de las características de los aprendices. En particular en la enseñanza en preescolar se caracteriza por ser el primer ciclo de la educación obligatoria, en la cual se plantea la utilización de prácticas adecuadas al desarrollo y aprendizaje de los niños.

En general, los programas de evaluación de la docencia de los profesores de educación básica incorporan información de tres fuentes principales: 1) del aprendizaje de los estudiantes al final del curso; 2) evaluaciones hechas por pares académicos y directivos; y 3) autoevaluaciones, autorreportes y portafolios. La elección de una estrategia particular de evaluación depende en gran medida de los propósitos de la evaluación y de las condiciones institucionales de aplicación. La evaluación por pares académicos, como estrategia, ha sido ampliamente utilizada con fines sumativos y formativos (Millis, 2006). Aunque en la actualidad se ha privilegiado su uso en contextos orientados a la mejora de la actividad docente.

La presente investigación tiene como propósito diseñar y aportar evidencias de validez de un instrumento de observación en el aula para evaluar la práctica de las docentes de educación preescolar en México.

En el plano internacional existen dos instrumentos de observación ampliamente utilizados en educación preescolar. El primero es el *Early Chilhood Enviromennt Rating Scale* (ECERS) cuyo propósito es medir la calidad de los servicios educativos en edad preescolar, es un instrumento estandarizado de observación, los referentes de evaluación que le subyacen están vinculados con el desarrollo de prácticas adecuadas para este nivel educativo (Harms, Clifford y Debby, 1988).

El segundo es el *Classrom Assessment Scoring System* (CLASS); es un protocolo estandarizado de observación de la práctica docente, su propósito es medir la calidad de las interacciones del docente relacionadas con un mayor aprendizaje de los alumnos, el cual ha sido utilizado principalmente en EUA (Pianta y Hamre, 2009). Ambos instrumentos se caracterizan por tomar como marco de referencia la investigación sobre el desarrollo y aprendizaje infantil, no están referidos a un currículo en particular.

En México, Myers, Martínez y Linares (2003) desarrollaron un instrumento de observación similar al ECERS que evalúa el ambiente educativo y tiene como referente el concepto de calidad de la educación preescolar, cuyo fin último es el aprendizaje para una cultura democrática. El instrumento se ha utilizado principalmente en proyectos de investigación. No obstante, aunque se haya utilizado en distintos momentos no se han presentado las propiedades métricas de la escala.

La Secretaria de Educación Pública (SEP) ha realizado investigaciones de la enseñanza en el marco de la Reforma a la Educación Preescolar. El equipo responsable en la Dirección de Desarrollo Curricular para Preescolar emprendió un conjunto de acciones para promover el conocimiento y análisis de –entre otros aspectos- la práctica pedagógica en las aulas del país. Derivado de estas acciones se realizó un diagnóstico de la situación del preescolar en México (SEP, 2006) el cual sirvió principalmente para

identificar avancen en la implementación del Programa de Educación Preescolar 2004 (PEP 2004) y establecer estrategias de mejora.

Pedroza, Álvarez y Jiménez (2013), en el marco de una evaluación de la implementación del currículo de preescolar [PEP 2004], desarrollaron una bitácora para evaluar de manera estandarizada las prácticas docentes. Para ello se utilizó un auto registro con preguntas abiertas para identificar lo que hacían las educadoras en tres jornadas escolares, el cual fue codificado con un conjunto de rúbricas. Se identificaron los propósitos educativos, las características del enfoque pedagógico y la demanda cognitiva de las actividades.

Una evaluación alineada al currículo y al marco normativo de la evaluación docente es fundamental, puesto que la evaluación orienta la acción educativa. Darling-Hamond (2012) señaló que para que la evaluación docente tenga un efecto positivo en el aprendizaje de los niños tendría que partir de una estrategia sistémica de evaluación, compuesta por: estándares comunes de evaluación; evaluaciones de desempeño basados en estos estándares; sistemas locales de evaluación inspirados en los mismos estándares para evaluar la calidad de la enseñanza *in situ*; estructuras de apoyo, evaluadores y mentores para docentes que requieran ayuda adicional; y oportunidades de desarrollo profesional.

A partir de la revisión de los antecedentes, es claro que en México no existe un instrumento en preescolar que evalúe la práctica docente mediante observación y análisis de evidencias tomando como referente el marco normativo. El instrumento que se presenta en este artículo se diseñó con la particularidad de estar alineado al currículo y al marco normativo para evaluar a los docentes de educación preescolar en México. Por una parte el instrumento retoma el Perfil, Parámetros e Indicadores [PPI] (SEP, 2015), que son los estándares para evaluar a los docentes, y por otra el Programa de Educación Preescolar 2011 [PEP 2011] (SEP, 2011), puesto que en este se establecen los principios pedagógicos que guían la práctica de los docentes de este nivel educativo.

2. Método

En este apartado se describe el proceso seguido para en el desarrollo de un instrumento de evaluación de la práctica docente en la educación preescolar.

2.1. Participantes

Se contó con un total de 28 participantes, incluyendo a los participantes en el diseño del instrumento como a los participantes en la muestra, a saber: ocho integrantes del comité de jueces de rúbricas, una educadora experta quien participó en la elaboración de las rúbricas y 19 educadoras a quienes se aplicó el instrumento. Las características de los participantes son las siguientes:

El comité de jueceo: se conformó por profesionales con experiencia en algunos de los siguientes aspectos: diseño del currículo nacional, evaluación en educación preescolar y asesoría técnico pedagógica al profesorado. Su media de edad es 46 y los años de experiencia 24 años.

La educadora experta: participó en la elaboración de las rúbricas, es jubilada con 28 años de servicio; 24 frente a grupo y con 15 de docente en educación normal y 4 de asesor técnico pedagógico del nivel preescolar.

Educadoras observadas: el grupo se conformó por 19 docentes con formación específica para la docencia en el nivel preescolar; todas del sexo femenino. Diez fueron educadoras en formación en su primer año frente a grupo y las nueve restante con más de seis años de experiencia. Centra (1993) identifica a los docentes en la etapa de noveles (0-3 años) hasta de mayor experiencia (más de 8 años). Las docentes laboraban en escuelas públicas ubicadas en zonas urbanas de alta marginación.

2.2. Materiales

- ✓ Dos videocámaras, equipadas con tripié y micrófonos inalámbricos de solapa.
- ✓ Equipo de cómputo (hardware y software) para procesamiento, almacenamiento y acceso en línea al material audiovisual (Google drive) proporcionado por la Universidad Autónoma de Baja California.
- ✓ Cuadernillo del instrumento y hojas de registro.

2.3. Procedimiento

Este estudio comprendió tres etapas: desarrollo del instrumento, recolección de la información y análisis de los resultados. Cada una de ellas se describe a continuación.

2.3.1. Etapa 1: Desarrollo del instrumento

El desarrollo del instrumento tuvo las siguientes fases: operacionalización, elaboración de rúbricas, jueceo de las rúbricas por el comité de validación, prueba en campo y elaboración de manuales.

- ✓ *Fase 1. Operacionalización.* La tarea consistió en definir constitutivamente el constructo que mide el instrumento: la enseñanza en educación preescolar. Se retomaron los postulados de Shulman (2005), y se dividió el constructo en tres dimensiones: planeación, intervención educativa y evaluación. A continuación se presentan las definiciones de las mismas.

  Planeación. Incluye dos procesos tal como los señaló Shulman (2005), el de comprensión y el de transformación. La comprensión es donde el docente conoce y comprende críticamente un conjunto de ideas que van a enseñarse. Además de la comprensión del contenido a enseñar, es fundamental que el docente comprenda los objetivos educativos, que si bien pueden partir de un texto o currículo, estos tiene que adaptarse o modificarse en función del grupo de educandos.

  El segundo proceso es la transformación, en este las ideas comprendidas deben ser modificadas para poder enseñarlas. En general implican varios pasos: preparación (materiales, textos, etc.); representación de las ideas a nuevas formas como analogías o metáforas; selección de un método didáctico; y adecuación de estas adaptaciones a las características generales del grupo y a características específicas de cada niño de la clase.

  Para el PEP 2011, la planeación de la intervención educativa es entendida como "un conjunto de supuestos fundamentados que la educadora considera pertinentes y viables para que niñas y niños avancen en su proceso de aprendizaje" (SEP, 2011: 25). Asimismo, señala que la planeación es indispensable para que el docente realice una práctica eficaz, ya que le permite definir la intención educativa, prever los recursos didácticos, la organización del

grupo e identificar los criterios para evaluar el proceso de aprendizaje en los alumnos que conforman su escolar. Es claro que el programa visualiza este proceso, al igual como lo señala la literatura, como una tarea reflexiva que implica distintos elementos a considerar para poder realizarse adecuadamente.

Intervención educativa. Es la actividad que comprende el desempeño observable del docente en el aula. Incluye distintas acciones de la didáctica como: la organización y manejo del grupo; las explicaciones; la asignación de trabajo o tareas y su revisión; el monitoreo a los alumnos; y la interacción por medio de preguntas, reacciones, elogios o críticas Shulman (2005). A esta fase de la enseñanza otros autores denominan como conducción del proceso de enseñanza - aprendizaje (García, Loredo, Luna, Rueda, 2008). Los distintos elementos de la intervención educativa se organizaron en dos grandes aspectos: a) principios pedagógicos del programa y b) ambiente de aprendizaje.

La evaluación. Es un proceso que refiere a la verificación de la comprensión en los alumnos, incluye las distintas actividades mediante las cuales los docentes identifican el progreso de sus alumnos y reconocen el logro de las metas de aprendizaje, como verificar o monitorear la comprensión de los alumnos durante el proceso de enseñanza, retroalimentar la comprensión de los alumnos al finalizar las situaciones didácticas o unidades de aprendizaje. En otros niveles educativos, este proceso también incluye una valoración para emitir una calificación. Para la comprobación de los aprendizajes en los niños se requiere que el docente comprenda el contenido a enseñar y los procesos de aprendizaje del niño, es decir, las formas de comprensión y transformación descritas anteriormente. Otra orientación de la evaluación, es hacia el propio docente y su desempeño logrado, lo que conduce a otro proceso de la enseñanza, que es el de reflexión sobre la práctica.

Posteriormente se definió operacionalmente el constructo, para ello se generó una tabla donde se ordenaron los elementos del más abstracto al más concreto: dimensiones, subdimensiones e indicadores. Esta tabla de contenidos se acompañó de un documento de fundamentación, donde se describieron los contenidos a evaluar y se identificó su relación con el PEP 2011.

- ✓ *Fase 2. Elaboración de rúbricas.* En el diseño de las rúbricas se siguió una metodología de elaboración colectiva, a partir de la deliberación colegiada de los participantes (Jornet, González, Suárez y Perales, 2011; Pedroza, Vilchis, Álvarez, López y García, 2013). El proceso de construcción de las rúbricas se realizó de acuerdo con el siguiente orden: descripción del nivel 3 que corresponde a lo esperado por el programa; luego se describe el nivel 1, características de una práctica inadecuada; después del nivel 2, en donde se mezclan elementos del nivel 1 y 3; y al final el nivel más avanzado, el nivel 4, donde se incluye un *plus* respecto a lo que señala el programa. En la elaboración, se consideraron los criterios de calidad de las rúbricas descritos por Arter (2010). La elaboración tomó aproximadamente 90 horas de trabajo. El producto de esta fase es una primera versión de las rúbricas.
- ✓ *Fase 3. Jueceo del instrumento.* Para esta fase se conformó el Comité de Jueceo, el cual, durante dos sesiones colectivas, estuvo encargado de hacer la valoración de la primera versión de las rúbricas en función de su relevancia y suficiencia con el

PEP 2011. Para esto se siguió un procedimiento donde se privilegió llegar a consensos sobre las modificaciones a realizar en el instrumento.

- ✓ *Fase 4. Piloteo de las rúbricas.* En esta fase se realizaron dos acciones, en primer lugar se realizó una prueba del instrumento a partir de evidencias como las planeaciones de docentes, expedientes de los niños y videos de la práctica docente. Se analizaron 12 planes de trabajo, tres videos y ocho expedientes. Se siguió un proceso interactivo, donde se evaluaba la evidencia y se hacían modificaciones al instrumento. En la segunda acción se llevó a cabo una prueba en campo, la cual consistió en la observación de cuatro jornadas escolares completas a distintos docentes. Esto permitió examinar el funcionamiento del instrumento en condiciones de aplicación y realizar los ajustes pertinentes del mismo. La etapa culminó con la versión final del instrumento y su manual de uso.

2.3.2. Etapa 2: Recolección de información y retroalimentación a educadoras

- ✓ *Fase 1. La conformación de la muestra.* La conformación de la muestra de educadoras se dio a partir de aquellas que dieron acceso a que fueran grabadas en video. En el diseño del proyecto inicialmente se contempló que fueran 20 educadoras, sin embargo, de una no fue posible completar todas las observaciones. La aplicación del instrumento se realizó a 19 educadoras (tabla 1).

Tabla 1. Características de la muestra de educadoras observadas

	N
Jardines de niños	3
Educadoras	19
Sesiones observadas por educadora	3
Codificaciones por sesión	2
Total de registros de observación	114

Fuente: Elaboración propia.

Se estableció un acuerdo en el que los investigadores se comprometieron a devolver una retroalimentación para cada una de las observaciones y poner a disposición de las docentes una copia electrónica de los videos. Además, se contó con un convenio en el que las educadoras dan su consentimiento para la grabación de su clase y en el que se establecen los usos de la información recabada, los cuales se rigen bajo la Ley Federal de Protección de Datos Personales en Posesión de Particulares (DOF, 2010). La información proporcionada es utilizada bajo los principios de confidencialidad, consentimiento, información, calidad, finalidad, lealtad, proporcionalidad y responsabilidad.

- ✓ *Fase 2. Calibración de la observación.* Dado que los dos observadores participaron en la elaboración de las rúbricas, no se requirió un proceso de capacitación para el uso del instrumento. En vez de ello se realizó una codificación colegiada a partir de videos, lo que permitió unificar los criterios de valoración.
- ✓ *Fase 3. Trabajo de campo.* La recolección de información se llevó a cabo en un periodo de tres meses. Los observadores estuvieron encargado de recabar las planeaciones didácticas y grabar la intervención de cada docente durante toda la jornada escolar, exceptuando el recreo y clases que la docente titular no dirigía

(música, educación física, etc.). Cada educadora fue grabada en video durante tres sesiones de trabajo, y cada ocasión se calificó dos veces, una *in situ* y otra a partir del video. De esta forma se cuenta con seis medidas de desempeño de cada educadora (ver Tabla 1).

- ✓ *Fase 4. Retroalimentación a los evaluados.* Los observadores estuvieron encargados de entregar un reporte de resultados y de dar retroalimentación individual a las educadoras al finalizar los ciclos de grabación. Además, los videos se pusieron a disposición de las docentes a través de un acceso personalizado en línea.

2.3.3. Etapa 3. Análisis de la información

En esta etapa se sistematizó la información para conocer la estructura y confiabilidad del instrumento, así como la precisión de las estrategias de recolección de información. La información recabada se analizó en tres fases:

- ✓ *Fase 1. Identificación de la dimensionalidad del instrumento.* Los ítems del instrumento se sometieron a un análisis con el Modelo de Crédito Parcial de Rasch Master (Masters, 1982), con la finalidad de identificar ítems anómalos en su funcionamiento, este modelo es utilizado para variables ordinales. Si bien se reconoce que el modelo Rasch para ítems dicotómicos es sensible al tamaño de la muestra, Smith, Fallowfield, Velikova y Sharpe (2008) encontraron que los modelos politómicos de Rasch no presentan la misma sensibilidad, aunque se identificó mayor estabilidad en muestras mayores a 200 sujetos.

  Posteriormente se realizó un análisis factorial exploratorio (AFE) para probar la estructura subyacente a partir de la evidencia empírica. Dodou y Wieringa (2009 citados en Frías-Navarro, Pascual-Soler, 2012) señalaron que un N=50 es un valor mínimo razonable para este tipo de análisis, en este estudio tuvieron 114 observaciones. Para el AFE se utilizó un módulo del paquete R denominado *R Factor* (Basto y Pereira, 2012), el cual pone a disposición distintas técnicas de extracción, rotación y selección del número de factores. Debido a que la métrica de las rúbricas es ordinal, se decidió utilizar el análisis factorial a partir de matrices policóricas. El método de extracción fue Factorización de Ejes Principales. Se hizo una rotación oblicua (Quartimin) ya que es más pertinente para variables sociales donde se espera que los factores tengan algún grado de asociación (Basto y Pereira, 2012). Para la selección del número de factores se valoraron cinco métodos: Regla de Kaiser, Análisis de Paralelo, Coordinación Óptima, Factor de aceleración y Prueba de Velicer. Asimismo, se estimaron varios coeficientes de confiabilidad como: alfa de Cronbach, coeficiente alfa ordinal y theta de Armor´s.

- ✓ *Fase 2. Confiabilidad de las dimensiones.* Se calcularon cuatros coeficientes de confiabilidad: alfa de Cronbach, el coeficiente alfa ordinal, Theta de Armor y Coeficiente Theta ordinal. Los tres últimos coeficientes son más apropiados para una métrica ordinal.

- ✓ *Fase 2. Confiabilidad de la observación.* Se realizaron tres análisis para dar cuenta del grado de acuerdo que presentan los evaluadores: porcentaje de acuerdo entre los jueces, Kappa de Cohen y coeficientes de determinación y generalizabilidad. En primer lugar se calculó el porcentaje de acuerdo exacto entre los dos observadores. En un segundo momento se estimaron los coeficientes Kappa de

Cohen en su versión cuadrático ponderado. Por último, a partir de la teoría de la generalizabilidad se estimó la confiabilidad entre los observadores y entre las distintas sesiones de observación, así como las fuentes de varianza principales. El análisis de generalizabilidad ofrece una mejor medición de los errores que otras formas de análisis porque permite identificar las variaciones atribuidas a distintas facetas de medida como juez, ocasión, ítems, y las interacciones entre estas de forma simultánea (Shavelson y Webb, 1991).

3. Resultados

3.1. Elaboración de rúbricas

El propósito del instrumento es evaluar la práctica pedagógica para ofrecer una retroalimentación al desempeño del docente. Enseguida se presentan algunas de sus características:

- ✓ *Población objetivo.* Docentes de educación preescolar.
- ✓ *Uso.* Evaluación formativa de docentes e investigación.
- ✓ *Tipo de instrumento.* Protocolo de observación, compuesto por un conjunto de rúbricas que especifican los criterios de calidad de la enseñanza para distintos niveles de desempeño.
- ✓ *Forma de administración*: mediante la observación, ya sea in situ, o por medio de grabaciones en video de unidades didácticas completas.
- ✓ *Usuarios:* agentes del sistema educativo encargados de hacer una supervisión o acompañamiento de la práctica pedagógica como supervisores, asesores técnico pedagógicos ATP's, directores y colegas docentes, así como observadores externos con conocimiento del PEP 2011.
- ✓ *Perfil de los evaluadores*: conocimiento del currículo de preescolar y experiencia en la observación de la práctica docente. Asimismo, acreditar un proceso de capacitación, en el cual deberán mostrar un grado de acuerdo superior a 80% con los criterios del diseñador.
- ✓ *Calificación*: el puntaje del instrumento está determinado por las rúbricas, donde cada nivel de desempeño, corresponde a un puntaje distinto.
- ✓ *Devolución de resultados*: posterior a la calificación del instrumento se genera un reporte individualizado, donde además del puntaje de cada rúbrica se incluye la descripción del nivel de desempeño. La devolución de la información a las educadoras se da en entrevista presencial, privilegiando el diálogo entre el evaluador y el evaluado.

En la primera versión del instrumento se construyeron 38 matrices de valoración. Las matrices tienen cuatro niveles de calidad para cada rubro, salvo algunos casos donde es más pertinente tener tres niveles (tabla 2). Estos niveles reflejan un continuo de experiencia en la enseñanza, una progresión de niveles menos desarrollados a niveles más desarrollados. El nivel "insatisfactorio", el más bajo, representa una práctica inadecuada o errónea desde el PEP 2011; el nivel "en proceso", es una práctica que tiene rasgos de niveles más avanzados pero con mezcla de elementos del nivel insatisfactorio;

el nivel "competente" es una práctica que está de acuerdo a lo que señala el PEP 2011; y el nivel "experto" representa una práctica de un docente con amplia experiencia y que manifiesta una práctica ejemplar por su dominio y ejecución. El instrumento consta de una hoja de registro para una jornada de observación y un cuadernillo con las matrices de valoración. Enseguida se presentan sus características principales.

Tabla 2. Ejemplo de las matrices de valoración del instrumento

DIMENSIÓN	INSATISFACTORIO	EN PROCESO	COMPETENTE	EXPERTO
Intención educativa	Las intenciones educativas no están vinculadas con las competencias o aprendizajes esperados del programa.	Las intenciones educativas están vinculadas a las competencias del programa o con alguno de los aprendizajes esperados.	Las intenciones educativas están vinculadas a las competencias del programa y con alguno los aprendizajes esperados.	Las intenciones educativas están vinculadas a las competencias del programa y con algunos de los aprendizajes esperados. Además se identifican competencias o aprendizajes esperados que se favorecen indirectamente con la situación didáctica.
Conocimientos previos	La docente no recupera los conocimientos previos de los alumnos. No hace preguntas o actividades de indagación.	La docente recupera los conocimientos previos de los alumnos en forma parcial. Utiliza preguntas que no permiten indagar la intención educativa a trabajar.	La docente recupera los conocimientos previos de los alumnos y corresponden a la intención educativa a trabajar.	La docente recupera los conocimientos previos de los alumnos, corresponden a la intención educativa a trabajar y los vincula con la situación didáctica.
Promueve la interacción entre niños	La educadora no promueve la interacción entre los niños o les limita que se comuniquen entre ellos.	La educadora permite la interacción entre los niños. Existe un acomodo de los alumnos en pequeños grupos. Pero se designa sólo trabajos individuales, o de manera grupal todos tienen el mismo objetivo y producto.	La educadora promueve la interacción entre los niños. Existe una acomodo de los alumnos en pequeños grupos y trabajan colaborativamente (realizan un producto en equipo).	La educadora promueve la interacción colaborativa de los niños en distintas formas, por ejemplo: en parejas, en pequeños grupos, en trabajo grupal, sesiones plenarias.

Fuente: Elaboración propia.

3.2. Jueceo del instrumento

A partir de la retroalimentación realizada por el comité de jueceo se realizaron los siguientes cambios: diez rúbricas fueron aprobadas sin cambios, la mayoría tuvieron modificaciones y se eliminaron siete de ellas. La eliminación de las rúbricas se debió a que estas indagaban aspectos poco relevantes de la práctica docente o el contenido no

podía evaluarse adecuadamente mediante las evidencias: observación de tres sesiones, plan de clase y expedientes.

Las modificaciones de las rúbricas obedecieron principalmente a las razones: mejorar la descripción de los indicadores a observar, ya sea por una redacción ambigua o porque se incluían ejemplos que no eran apropiados; las descripciones, principalmente de los niveles más altos, creaban una imagen poco real de la práctica docente, se incluían elementos que se realizan con poca frecuencia en los jardines de niños; se modificaron dos rúbricas que no representaban de forma fiel lo que el PEP 2011 señala; otro grupo numeroso de cambios, se debió a observaciones de forma, que incluyeron cambios en la redacción, de una palabra por otra y cambios sintácticos-gramaticales. Cabe mencionar que no se sugirieron cambios a la estructura de la operacionalización.

Tabla 3. Cambios realizados en las rúbricas del instrumento por el comité de jueceo

DIMENSIÓN	PRESENTADAS A REVISIÓN	APROBADAS	MODIFICADAS	ELIMINADOS
Planeación	16	1	8	7
Intervención Principios pedagógicos	11	5	6	0
Ambiente de aprendizaje	7	1	6	0
Evaluación	4	0	4	0
Totales	38	7	24	7

Fuente: Elaboración propia.

3.3. Pilotaje del instrumento

El pilotaje del instrumento se realizó en cuatro aulas, esto permitió identificar elementos que no son factibles de observar durante una sesión de clase y precisar las descripciones de las rúbricas. A partir de la observación en campo se eliminó una rúbrica más, puesto que en una sesión de clase no se lograba identificar los aspectos a observar. Asimismo, se hicieron ajustes a once rúbricas. La estructura final del instrumento se presenta en la tabla 4.

El estudio piloto también sirvió para probar el proceso de recolección de la evidencia para la aplicación definitiva, como el plan de trabajo y la grabación de la práctica de la docente en video. El proceso de aplicación del instrumento es el siguiente:

1. *Obtención de la planeación de clase.* El observador solicita la planeación a la docente antes de iniciar la sesión. Se califican las primeras rúbricas correspondientes a la planeación, a excepción de las relacionadas con el conocimiento de los niños. En este paso también se fotografía la planeación.
2. *Observación de toda la jornada escolar*, durante este tiempo se graba en video la sesión y se califica las rúbricas del instrumento.
3. *Procesamiento y almacenamiento del material audiovisual.* Se organiza y almacena en medios electrónicos las evidencias obtenidas.
4. *Segunda codificación.* Se analiza la planeación didáctica y se observa la práctica docente por medio del video.
5. *Retroalimentación.* Se genera un reporte individualizados por sesión de observación y en entrevista con la docente se le ofrece una retroalimentación.

Tabla 4. Estructura del instrumento para su aplicación extensiva

PROCESO	DIMENSIÓN	SUB-DIMENSIONES
1. Planeación	1.1. Intención educativa	1.1a. Intención congruente con el PEP 2011 1.1b. Intención acorde a las necesidades de niños
	1.2. Conocimiento de los alumnos	1.1c. Intención clara
	1.3. Diseño de la situación de aprendizaje	1.3a. Actividad congruente con intenciones 1.3b. Actividad congruente con enfoque del campo 1.3c Plan para evaluación
	1.4. Demanda cognitiva	
2. Intervención educativa	2.1. Principios pedagógicos del programa	2.1. Recupera conocimientos previos 2.2. Promueve la interacción entre los niños 2.3. Fomenta el deseo por aprender 2.4a. Hace el contenido interesante (estrategias) 2.4b. Interés de los niños en la actividad 2.5. Promueve la participación y responsabilidad en el aprendizaje 2.6. Reglas 2.7. Demanda cognitiva de la actividad 2.8. Dominio del contenido 2.9. Congruencia con lo planeado
	2.2. Ambiente del aula para aprendizaje	2.10a. Comunicación cálida (verbal) 2.10b. Comunicación cálida (no verbal) 2.11. Consigna clara 2.12. Manejo de errores 2.13. Orden del grupo 2.14a. Uso del tiempo en las actividades 2.14b. Uso del tiempo en la jornada 2.15. Atención a la diversidad
3. Evaluación de la interacción en aula		3.1. Monitoreo 3.2. Retroalimentación 3.3. Reflexión sobre el proceso de aprendizaje 3.4. Reconocimiento

Fuente: Elaboración propia.

3.4. Análisis descriptivos

La distribución de los datos muestra que, en la gran mayoría de los casos, las docentes observadas se distribuyeron en todos los niveles definidos en cada una de las rúbricas de Planeación (tabla 5). Sin embargo, se observa que en dos de ellas, en la 1.1a y 1.1c, un alto porcentaje de las docentes se concentra en el nivel más alto de la rúbrica (experto); esto se relaciona con que se encontraron planes de trabajo muy exhaustivos entre las educadoras observadas, debido a que las directoras exigen a ese nivel de detalle las planeaciones. No obstante, esto no sucedió así en la muestra de planeaciones que se revisó en la construcción del instrumento.

Tabla 5. Distribución de los evaluados por nivel de dominio. Planeación

RÚBRICAS	PORCENTAJE DE EVALUADOS POR NIVEL				TOTAL
	Insatisfactorio	En proceso	Competente	Experto	
1.1a. Intención congruente con el PEP 2011	3.5%	2.6%	8.8%	**85.1%**	100.0%
1.1b. Intención congruente con necesidades de niños	5.3%	14.0%	33.3%	47.4%	100.0%
1.1c. Intención clara	2.6%	2.6%	2.6%	**92.1%**	100.0%
1.2. Conocimiento de los alumnos	4.4%	35.1%	34.2%	26.3%	100.0%
1.3a. Actividad congruente con intenciones	15.8%	32.5%	7.0%	44.7%	100.0%
1.3b. Actividad congruente con enfoque del campo	14.0%	40.4%	45.6%	N/A	100.0%
1.3c. Plan para evaluación	11.4%	39.5%	36.8%	12.3%	100.0%
1.4. Demanda cognitiva de planeación	10.5%	48.2%	28.1%	13.2%	100.0%

Fuente: Elaboración propia.

Una distribución similar se observa en las rúbricas de Intervención y Evaluación, los informantes se distribuyen, por lo general, en todos los niveles (Tabla 6). Sólo en cuatro de ellas se presenta una alta concentración de evaluados en uno de los niveles (2.2, 2.12, 2.15 y 2.9); destaca el caso de la rúbrica 2.12, en la cual los extremos están prácticamente vacíos, esto sugiere que sería factible colapsar, para este caso, los 4 niveles a sólo 2. Además, se identifican dos rúbricas en las que el nivel más alto está desierto (2.5 y 3.3), lo que invita a revisar la pertinencia de dicho nivel.

Tabla 6. Distribución de los evaluados por nivel de dominio. Intervención y Evaluación

RÚBRICAS	PORCENTAJE DE EVALUADOS POR NIVEL				TOTAL
	Insatisfactorio	En proceso	Competente	Experto	
2.1. Recupera conocimientos previos	13.2%	34.2%	24.6%	28.1%	100%
2.2. Promueve la interacción entre los niños	0.9%	**79.8%**	10.5%	8.8%	100%
2.3. Fomenta el deseo por aprender	7.9%	64.0%	28.1%	N/A	100%
2.4a. Hace el contenido interesante (estrategias)	1.8%	34.2%	36.0%	28.1%	100%
2.4b. Interés de los niños en la actividad	5.3%	21.1%	48.2%	25.4%	100%
2.5. Promueve la participación y responsabilidad en el aprendizaje	11.4%	60.5%	28.1%	**0.0%**	100%
2.6. Reglas	1.8%	29.8%	65.8%	2.6%	100%
2.7. Demanda cognitiva de la actividad	14.9%	44.7%	36.0%	4.4%	100%
2.8. Dominio del contenido	3.5%	14.9%	65.8%	15.8%	100%
2.9. Congruencia con lo planeado	8.3%	4.6%	**87.2%**	N/A	100%
2.10a. Comunicación cálida (verbal)	0.9%	34.2%	50.0%	14.9%	100%
2.10b. Comunicación cálida (no verbal)	0.0%	58.8%	39.5%	1.8%	100%
2.11. Consigna clara	0.9%	27.2%	47.4%	24.6%	100%
2.12. Manejo de errores	0.0%	**71.1%**	28.1%	0.9%	100%
2.13. Orden del grupo	7.0%	37.7%	51.8%	3.5%	100%
2.14a. Uso del tiempo en las actividades	3.5%	42.1%	44.7%	9.6%	100%
2.14b. Uso del tiempo en la jornada	5.3%	42.1%	52.6%	N/A	100%
2.15. Atención a la diversidad	9.7%	13.3%	**74.3%**	2.7%	100%
3.1. Monitoreo	5.3%	43.9%	29.8%	21.1%	100%
3.2. Retroalimentación	20.2%	53.5%	25.4%	0.9%	100%
3.3. Reflexión sobre el proceso de aprendizaje	57.9%	24.6%	17.5%	**0.0%**	100%
3.4. Reconocimiento	30.7%	43.0%	18.4%	7.9%	100%

Fuente: Elaboración propia.

3.5. Análisis según el modelo de Rasch Masters

El primer análisis del funcionamiento de las rúbricas del instrumento se realizó con Rasch. Los resultados indican que las mayoría de las rúbricas se ajustan de forma adecuada al modelo. Es decir que los estadígrafos INFIT y OUTFIT se encuentran entre un rango de .7 y 1.3, o la correlación punto biserial es mayor a .40 (Linacre, 2007). Sólo cuatro de ellas están fuera de estos parámetros. En la tabla 7 se puede identificar que el nivel de dificultad que van desde – 2.2 a 3.35 en escala *logit*.

Tabla 7. Estadígrafos de ajuste del modelo Rasch para las rúbricas del instrumento

Ítems	Dificultad	INFIT	OUTFIT	Correlación Punto Biserial	Discrimina
1.1a. Intención congruente con el PEP 2011	-1.95	0.78	**1.52**	0.48	1.02
1.1b. Intención congruente con necesidades de niños	-1.04	0.94	0.92	0.63	1.06
1.1c. Intención clara	-2.2	0.72	1.38	0.42	1.03
1.2. Conocimiento de los alumnos	-0.52	0.95	0.93	0.64	1.08
1.3a. Actividad congruente con intenciones	-0.21	1.18	1.09	0.64	0.89
1.3b. Actividad congruente con enfoque del campo	-0.56	0.84	0.78	0.65	1.24
1.3c. Plan para evaluación	0.35	1.2	1.19	0.52	0.76
1.4. Demanda cognitiva de planeación	0.36	1.4	**1.53**	0.4	0.52
2.1. Recupera conocimientos previos	-0.02	1.15	1.17	0.59	0.84
2.2. Promueve la interacción entre los niños	-0.17	0.81	0.62	0.52	1.13
2.3. Fomenta el deseo por aprender	1.14	0.86	0.83	0.61	1.16
2.4a. Hace el contenido interesante (estrategias)	-0.93	0.95	0.94	0.62	1.06
2.4b. Interés de los niños en la actividad	-0.58	0.98	0.99	0.59	1
2.5. Promueve la participación y responsabilidad en el aprendizaje	-0.27	0.83	0.82	0.62	1.22
2.6. Reglas	0.00	0.85	0.82	0.58	1.11
2.7. Demanda cognitiva de la actividad	0.96	1.10	1.10	0.52	0.88
2.8. Dominio del contenido	-0.60	1.35	1.35	**0.33**	0.73
2.9. Congruencia con lo planeado	-1.77	**2**	**4.84**	**0.14**	0.54
2.10a. Comunicación cálida (verbal)	-0.88	0.97	0.94	0.57	1.05
2.10b. Comunicación cálida (no verbal)	2.76	0.96	0.84	0.51	1.09
2.11. Consigna clara	-1.20	1.21	1.23	0.45	0.69
2.12. Manejo de errores	3.35	0.82	0.63	0.56	1.25
2.13. Orden del grupo	0.56	0.82	0.82	0.65	1.18
2.14a. Uso del tiempo en las actividades	-0.08	0.74	0.74	0.71	1.29
2.14b. Uso del tiempo en la jornada	-1.32	0.97	0.9	0.54	1.06
2.15. Atención a la diversidad	0.57	1.32	1.31	0.42	0.88
3.1. Monitoreo	-0.24	0.83	0.76	0.69	1.2
3.2. Retroalimentación	1.82	0.86	0.86	0.62	1.16
3.3. Reflexión sobre el proceso de aprendizaje	1.42	0.83	0.86	0.61	1.16
3.4. Reconocimiento	1.25	1.02	0.98	0.60	1.01

Fuente: Elaboración propia.

Los ítems que se salen de los parámetros de funcionamiento adecuado se excluyeron para los análisis subsecuentes. Entre ellos podemos destacar el primero del instrumento, el alto valor de outfit indica que las docentes que no tenían la habilidad para puntuar alto en la rúbrica, sí lo hicieron.

3.6. Dimensionalidad del instrumento

La tTabla 8 presenta la estructura factorial del instrumento, en ella se puede apreciar que se conformaron tres factores. El primer factor está relacionado con el Ambiente de aprendizaje y las prácticas de evaluación, el segundo con la Planeación de la enseñanza y el tercero con los Principios pedagógicos del programa. Esto hace una diferencia respecto a la estructura teórica planteada, pues se consideraba a la evaluación como un factor independiente.

Tabla 8. Cargas factoriales de los ítems en los tres factores rotados

ÍTEMS	FACTORES		
	1	2	3
1.1b. Intención congruente con necesidades de niños	.227	**.694**	.018
1.2. Conocimiento de los alumnos	.153	**.727**	.077
1.3a. Actividad congruente con intenciones	.233	**.869**	-.115
1.3b. Actividad congruente con enfoque del campo	.210	**.747**	.034
1.3c. Plan para evaluación	-.026	**.692**	.082
2.7. Demanda cognitiva de la actividad	-.190	**.630**	.377
2.2. Promueve la interacción entre los niños	.247	.154	**.566**
2.3. Fomenta el deseo por aprender	.130	.168	**.636**
2.4a. Hace el contenido interesante (estrategias)	-.112	.304	**.780**
2.4b. Interés de los niños en la actividad	.156	.006	**.672**
2.5. Promueve la participación y responsabilidad en el aprendizaje	.217	.068	**.629**
2.10b. Comunicación cálida (no verbal)	.338	.025	**.438**
2.14a. Uso del tiempo en las actividades	.344	.250	**.423**
2.1. Recupera conocimientos previos	**.639**	.115	.029
2.6. Reglas	**.712**	.151	-.092
2.10a. Comunicación cálida (verbal)	**.612**	.026	.108
2.11. Consigna clara	**.422**	-.098	.283
2.12. Manejo de errores	**.717**	.195	-.029
2.13. Orden del grupo	**.754**	.013	.110
2.14b. Uso del tiempo en la jornada	**.491**	-.071	.372
2.15. Atención a la diversidad	**.666**	.136	-.137
3.1. Monitoreo	**.860**	-.051	.070
3.2. Retroalimentación	**.612**	-.020	.260
3.3. Reflexión sobre el proceso de aprendizaje	**.599**	.223	.045
3.4. Reconocimiento	**.660**	-.278	.416

Fuente: Elaboración propia.

La Tabla 9 muestra los porcentajes de varianza a partir del Análisis Factorial Exploratorio. En ella se puede apreciar que la solución factorial final explica el 66% de la varianza, el factor de Evaluación y ambientes de aprendizaje obtuvo el porcentaje superior con 50%, seguido de Planeación con 9.8 % y Principios pedagógicos con 6 %.

Tabla 9. Porcentaje de varianza explicada y autovalores del instrumento

FACTOR	AUTOVALORES	% OF VARIANZA	ACUMULADA %
Evaluación y ambientes de aprendizaje	12.536	50.143	50.143
Planeación	2.462	9.849	59.991
Principios pedagógicos	1.555	6.220	66.211

Fuente: Elaboración propia.

La solución de tres factores se determinó a partir de la comparación de varios métodos, como se aprecia en la Figura 1. La regla de Kaiser (*eigenvalues*) retiene cinco factores, el análisis de Paralelo establece el primer punto de corte en tres factores, al igual que el

análisis de Coordinación Óptima, la prueba de Velicer´s arrojó un resultado de dos factores y el Factor de aceleración de uno (ver figura 1).

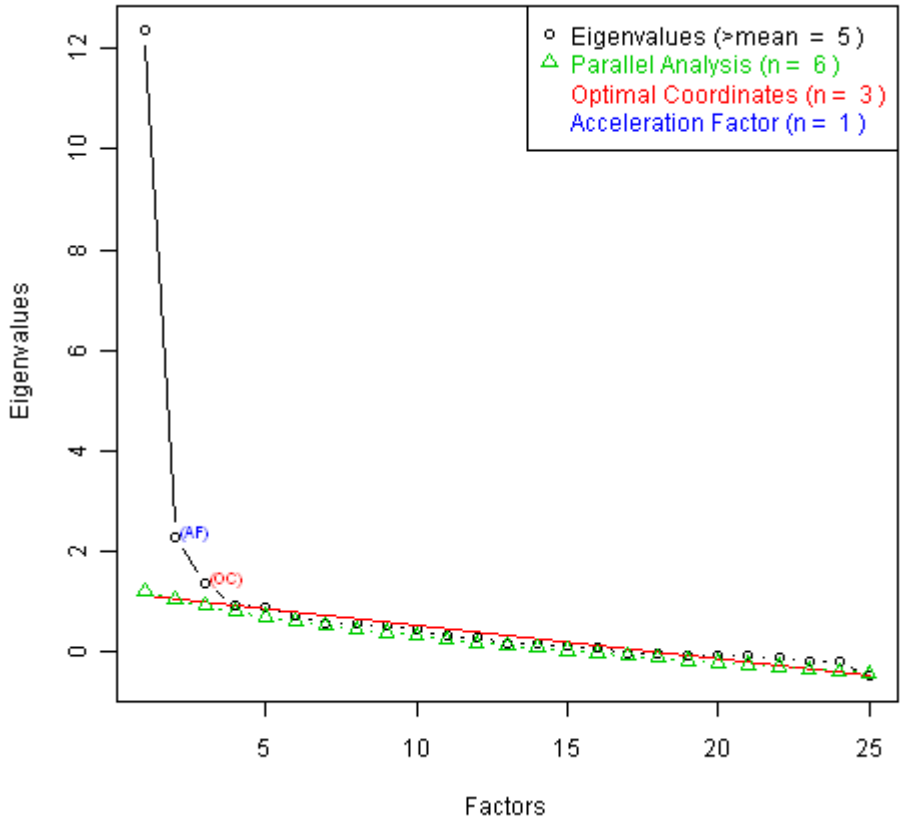


Figura 1. Estimación del número de factores a partir de cuatro métodos
Fuente: Elaboración propia.

Los métodos aproximaron a una selección de dos o tres factores. A partir de los estadísticos de bondad de ajuste de los modelos se identificó un mejor ajuste para la solución de tres factores. El índice de bondad de ajuste Goodness of Fit Index [GFI] tuvo un valor de .739 (valores cercanos al uno son considerados muy buenos), y el RMSR (Root Mean Square Residual) obtuvo un valor de .063 (valores menores a .05 son considerados excelentes).

3.7. Confiabilidad de las dimensiones

Se calculó la confiabilidad de los tres factores, mostrando valores cercanos al .90, lo que significa una consistencia alta para los ítems seleccionados. En la Tabla 10, además del usual alfa de Cronbach, se presenta el coeficiente alfa ordina. Además se presentan los valores de la confiabilidad utilizando dos estadísticos específicos para variables de tipo ordinal: Theta de Armor y Coeficiente Theta ordinal. En todos los coeficientes se muestran valores adecuados, mayores a 0.85.

Tabla 10. Coeficientes de confiabilidad para los factores del modelo

	FACTOR 1	FACTOR 2	FACTOR 3
Alfa de Cronbach	.893	.882	.852
Coeficiente alfa ordinal	.932	.918	.915
Theta de Armor	.896	.884	.855
Coeficiente theta ordinal	.933	.920	.915

Fuente: Elaboración propia.

3.8. Confiabilidad de la observación

En este apartado se presentan los resultados de la confiabilidad de la aplicación del instrumento. La Tabla 11 contiene el porcentaje de acuerdo exacto entre los jueces y el coeficiente de concordancia Kappa de Cohen en su versión cuadrática ponderada. Asimismo, se presenta el porcentaje de varianza atribuida a distintas facetas de error.

Tabla 11. Estadísticos sobre la confiabilidad de la codificación

ÍTEM	Porcentaje de acuerdo	Kappa cuadrática ponderada	PORCENTAJE DE VARIANZA							
			Educadora (E)	Ocasión (O)	Codificación (C)	E*O	E*C	O*C	E*O*C	Total
1.1a.	91	.84	38	0	0	47	2	1	12	100
1.1b.	68	.50	23	3	0	24	0	0	50	100
1.1c.	89	.54	25	1	0	29	10	0	36	100
1.2.	70	.70	31	12	0	28	1	0	28	100
1.3a.	58	.64	15	6	0	44	0	0	36	100
1.3b.	72	.61	27	3	0	29	0	0	41	100
1.3c.	63	.67	3	0	0	60	0	0	36	100
1.4.	65	.61	5	0	0	56	1	3	35	100
2.1.	56	.58	32	0	1	25	0	2	39	100
2.2.	81	.62	13	5	2	42	0	0	38	100
2.3.	67	.48	19	2	0	18	0	4	58	100
2.4a.	61	.60	25	3	1	32	0	0	40	100
2.4b.	65	.70	28	0	0	42	2	0	27	100
2.5.	58	**.37**	25	9	2	4	13	0	46	100
2.6.	77	.62	34	0	1	28	5	0	32	100
2.7.	72	.68	12	7	1	46	0	0	35	100
2.8.	70	.55	0	0	0	57	10	3	29	100
2.9.	89	.70	6	2	1	61	0	0	30	100
2.10a.	65	.53	26	4	0	24	2	2	41	100
2.10b.	53	**.22**	11	0	12	12	8	3	55	100
2.11.	53	**.39**	18	0	1	24	1	4	53	100
2.12.	86	.46	24	2	0	20	2	3	48	100
2.13.	84	.52	35	0	0	16	0	1	48	100
2.14a.	79	.53	22	0	2	32	5	0	39	100
2.14b.	91	.66	43	2	1	21	4	0	29	100
2.15.	95	.57	26	1	0	31	4	0	39	100
3.1.	81	.72	54	2	0	14	0	0	29	100
3.2.	77	.43	33	3	0	5	0	5	54	100
3.3.	91	.65	15	10	1	42	0	0	32	100
3.4.	81	.59	24	0	0	33	0	3	40	100
Evaluación y ambiente de aprendizaje			57	1	0	22	0	4	16	100
Planeación			28	8	0	36	0	0	28	100
Principios pedagógicos			42	5	0	33	0	1	19	100

Fuente: Elaboración propia.

A partir de los resultados se puede identificar que el porcentaje de acuerdo entre los jueces fue alto, el rango oscila entre 53% y 91%. El estadístico Kappa muestra grados de acuerdo un poco menores, esto debido a que esta medida ajusta el grado de acuerdo considerando el azar al contestar la escala. La escala de valoración para el coeficiente Kappa de Landis y Koch (1977) establece los siguientes valores: pobre (menos de .01), leve (entre .01 - .20), aceptable (.21-.40), moderada (.41-.60), considerable (.61-.80) y casi perfecta (mayor a .81).

Los resultados muestran que sólo tres valores caen en el rango aceptable, 11 como aceptable y el resto como moderados. Adicionalmente, en la Tabla 11 se presenta la descomposición de las variaciones de las puntuaciones para el estudio G. Para este tipo de diseño existen siete tipos de fuentes de varianza: entre las educadoras (E), entre las ocasiones de observación (O), entre las codificaciones (C), entre las educadoras en las distintas ocasiones (E*O), entre las codificaciones con las educadoras (E*C), la codificación en las distintas ocasiones (O*C) y una variación no identificada (E*O*C). En esta tabla 11 se incluyen todos los ítems del instrumento y al final tres índices correspondientes a cada factor resultantes de los análisis previos.

A partir del análisis de descomposición de la varianza, se identifica que algunos ítems tienen un funcionamiento deficiente, esto coincide con los ya identificados por los análisis con Rasch y factorial exploratorio: 1.1c, 1.4, 2.8 y 2.9. Además se identifican que algunos de ellos tienen dificultades para su codificación esto debido a que muestran una alta variación en los puntajes de los jueces, en este caso encontramos al 2.10b. A diferencia del estadístico Kappa, con el análisis de descomposición de la varianza, la variación entre los observadores no parece ser importante, de hecho, para la mayoría de los reactivos, las diferencias se aproximan a cero.

Por otra parte, los resultados permiten identificar dos fuentes principales de variación: la varianza real de las educadoras (E) y las educadoras en las distintas ocasiones (E*O). Para el primer caso significa que las educadoras mostraron un desempeño muy distinto entre ellas. La variación de las educadoras en las distintas ocasiones fue la segunda más alta para la mayoría de ítems, lo que significa que una misma educadora puede variar mucho su desempeño cada vez que se le evalúa. Para algunos ítems, la proporción de varianza fue mayor que incluso la variación entre educadoras.

En la tabla 12 se muestran los coeficientes de generalizabilidad para el diseño propuesto en esta investigación, el cual incluye tres sesiones de observación por dos observadores. La confiabilidad es mayor en el factor de ambientes de aprendizaje, el de principios del PEP es el segundo más alto. Por su parte, el relacionado con la planeación tuvo el valor más bajo, lo que indica que el diseño propuesto no permite obtener una medición precisa de esta dimensión de la enseñanza.

Tabla 12. Coeficientes de generalizabilidad para los índices del instrumento

ÍNDICE	COEFICIENTE G
Planeación	.63
Principios PEP	.75
Evaluación y ambiente de aprendizaje	.85

Fuente: Elaboración propia.

4. Discusión y conclusiones

El propósito de este estudio fue desarrollar y aportar evidencias de validez de un instrumento para la evaluación de la enseñanza de las docentes de educación preescolar. Los resultados de este artículo mostraron que con algunas modificaciones, el instrumento desarrollado ofrece información relevante de la enseñanza que realizan las educadoras en el aula.

Un instrumento como el aquí desarrollado es de utilidad para el sistema educativo por varias razones: (a) Los principios pedagógicos que evalúa el instrumento son los mismos que las docentes deben implementar en el aula, por lo tanto, los resultados obtenidos son pertinentes para conocer en qué medida los docentes muestran el desempeño esperado; (b) Los criterios de evaluación del instrumento están alineados con el currículo y la evaluación del desempeño docente, por esto, podría convertirse en una herramienta útil para apoyar a las educadoras durante su formación en servicio; y (c), El manejo del instrumento es relativamente sencillo, dado que no requiere del conocimiento de aspectos conceptuales adicionales a los que el personal del sistema educativo maneja en el desempeño de sus funciones (sólo se requiere conocer el PEP 2011).

Además de la pertinencia del instrumento con la normatividad vigente, los análisis realizados mostraron que cuenta con propiedades psicométricas adecuadas. La mayoría de ítems presentaron valores apropiados en el análisis Rasch Masters. Esto significa que las rúbricas tuvieron un correcto funcionamiento. Los análisis también permitieron identificar que algunos de los ítems requieren modificaciones para que puedan medir adecuadamente el constructo.

Los resultados obtenidos permitieron identificar una estructura subyacente con tres factores, la estructura teórica propuesta incluía tres dimensiones pero con otra configuración. Permaneció la dimensión de Planeación, pero no se identificó una dimensión de evaluación como en un inicio se tenía contemplado. En su lugar emergieron dos factores uno relacionado con los Principios pedagógicos del programa y otro con la Evaluación y ambientes de aprendizaje. Estos resultados hacen pensar que desde la perspectiva de las docentes estos dos factores forman parte de un evento pedagógico que no es independiente uno del otro. Por otra parte, también se identificó un factor dominante en la estructura subyacente: Evaluación y ambiente de aprendizaje, puesto que retiene una gran proporción de la varianza, dejando al resto de factores con una contribución proporcionalmente reducida. Los estudios recientes de la estructura factorial para el instrumento CLASS se han ensayado otras estructuras factoriales con dos factores y un factor utilizando Análisis Factorial Confirmatorio. Estas dos estructuras muestran ajustes adecuados, aunque la estructura de tres factores muestra un ajuste ligeramente mejor. A la estructura de un solo factor se le ha denominado de "Enseñanza Efectiva" (Hamre et al. 2013). Estos resultados apuntan a seguir explorando la estructura que subyace a la práctica docente de las educadoras, incorporando muestras más grandes y otras técnicas analíticas.

Se reconoce como limitación del estudio la reducida muestra de docentes observados. Si bien se cumple con los requerimientos en cuanto al tamaño de muestra para utilizar el AFE, es deseable tener muestras más grandes para los análisis Rasch de Crédito Parcial. Futuras indagaciones con el instrumento tendrán que aplicarse a una muestra más

amplia de sujetos en aras de ofrecer datos más precisos sobre el funcionamiento del instrumento y los ítems en particular.

Otro de los hallazgos relevantes del estudio es que las docentes muestran desempeños muy distintos de una sesión a otra. Se identificó que la segunda mayor fuente de variaciones en la práctica pedagógica, solo después de la variación entre educadoras, se concentra en las diferentes ocasiones en que se observó a las educadoras. Estos resultados coinciden con lo encontrado por otros investigadores (Martinez, Borko y Stecher, 2012). A partir de ello se confirma que si se quiere observar la práctica docente de forma precisa, es necesario observar en varias ocasiones a las docentes, pues es alta la variación que existe entre las distintas ocasiones en que se observa.

Referencias

Arter, J. A. (2010). Scoring rubrics. En P. Peterson, E. Baker y B. McGaw (Eds.), *International encyclopedia of education* (pp. 123-139). Nueva York, NY: Elsevier.

Basto, M. y Pereira, J. M. (2012). An SPSS R-menu for ordinal factor analysis. *Journal of Statistical Software*, *46*(4), 1-29. doi:10.18637/jss.v046.i04

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco, CA: Jossey-Bass.

Darling-Hammond, L. (2012). Desarrollo de un enfoque sistémico para evaluar la docencia y fomentar una enseñanza eficaz. *Pensamiento Educativo: Revista de Investigación Educacional Latinoamericana*, *49*(2), 1-20. doi:10.7764/PEL.49.2.2012.1

Diario Oficial de la Federación [DOF], (2010). Ley Federal de Protección de Datos Personales en Posesión de Particulares. Recuperado de: http://www.diputados.gob.mx/LeyesBiblio/pdf/LFPDPPP.pdf

Frías-Navarro, D. y Pascual Soler, M. (2012). Prácticas del análisis factorial exploratorio (AFE) en la investigación sobre conducta del consumidor y marketing. *Suma Psicológica*, *19*(1), 47-58.

García, B., Loredo, J., Luna, E. y Rueda, M. (2008). Modelo de evaluación de competencias docentes para la educación media y superior. *Revista Iberoamericana de Evaluación Educativa*, *1*(3), 124-136.

Hamre, B., Pianta, R., Downer, J., DeCoster, J., Mashburn, A., Jones, S. y Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*(4), 461-487. doi:10.1086/669616

Harms, T., Clifford, R. y Debby C. (1988). *Early Childhood Environment Rating Scale, Revised Edition (ECERS-R).* Williston, VT: Teachers College Press.

Gordon, R. A., Hofer, K. G., Fujimoto, K. A., Risk, N., Kaestner, R. y Korenman, S. (2015). Identifying high-quality preschool programs: New evidence on the validity of the Early Childhood Environment Rating Scale–Revised (ECERS-R) in relation to school readiness goals. *Early Education and Development*, *26*(8), 1086-1110. doi:10.1080/10409289.2015.1036348

Jornet, J. M., González Such, J. y Suárez, J. M. y Perales, M. J. (2011). Diseño de procesos de evaluación de competencias: Consideraciones acerca de los estándares en el dominio de las competencias. *Revista Bordón*, *63*(1), 125-145.

Landis, J. R. y Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. doi:10.2307/2529310

Linacre, J. M. (2007). *A user's guide to Winsteps Ministeps: Rasch-model computer programs*. Chicago, IL: Electronic Publication.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/bf02296272

Martínez, J. F., Borko, H. y Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: Lessons learned from two validation studies. *Journal of Research in Science Teaching*, *49*(1), 38-67. doi:10.1002/tea.20447

Millis, B. J. (2006). Peer observations as a catalyst for faculty development. En P. Seldin (Ed.), *Evaluating faculty performance. A practical guide to assessing teaching, research, and service* (pp. 82-95). San Francisco, CA: Anker Publishing Compay.

Myers, R., Martínez J. F. & Linares M.E., (2003). *En Búsqueda de la Calidad Educativa en Centros Preescolares*. Informe presentado a la Secretaría de Educación Pública. Dirección General de Investigación Educativa. Hacia una Cultura Democrática, A.C. México, D.F. Documento no publicado.

Pedroza, L. H., Álvarez, A. C. y Jiménez, A. B. (2013). *La implementación del PEP 2004 en las aulas*. En INEE (Eds.), *Prácticas pedagógicas y desarrollo profesional docente en Preescolar* (pp. 15-57). Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Pedroza, L. H., Vilchis, J. E., Álvarez, A. C., López, A. Y. y García, M. A. (2013). *Prácticas pedagógicas y desarrollo profesional docente en preescolar. Reporte técnico*. Ciudad de México: Instituto Nacional para la Evaluación de la educación. Recuperado de http://publicaciones.inee.edu.mx/buscadorPub//P1/D/240/P1D240_13E13.pdf

Pianta, R. C. y Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. doi:10.3102/0013189x09332374

Secretaría de Educación Pública (SEP). (2006). *La implementación de la reforma curricular en la educación preescolar: Orientaciones para fortalecer el proceso en las entidades federativas. Programa de Renovación Curricular y Pedagógica en la Educación Preescolar*. Ciudad de México: Autor.

Secretaría de Educación Pública (SEP). (2011). *Programa de Educación Preescolar 2011. Guía para la educadora*. Ciudad de México: Autor.

Secretaría de Educación Pública (SEP). (2015). *Evaluación del desempeño docente. Ciclo escolar 2015-2016. Perfil, parámetros e indicadores para docentes*. Ciudad de México: Autor.

Shavelson, R. J. y Webb, N. M. (1991). *A primer on generalizability theory*. Thousand Oaks, CA: Sage Publications.

Shulman, L. (2005). Conocimiento y enseñanza: Fundamentos de la nueva Reforma. *Revista de Currículum y Formación del Profesorado*, *9*(2), 1-30.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G. y Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*(1), 1-18. doi:10.1186/1471-2288-8-33.

Breve CV de los autores

Luis Horacio Pedroza Zúñiga

Estudiante de doctorado del Instituto de Investigación y Desarrollo Educativo, Universidad Autónoma de Baja California, México. ORCID ID: 0000-0002-5256-2967. Email: horaciopedroza@hotmail.com

Edna Luna Serrano

Investigadora del Instituto de Investigación y Desarrollo Educativo, Universidad Autónoma de Baja California, México. ORCID ID: 0000-003-1496-548X. Email: eluna@uabc.edu.mx

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 131-152.
https://doi.org/10.15366/riee2017.10.1.007

Adaptación de un Instrumento para la Medición de la Convivencia Escolar en Escuelas de Educación Secundaria de México

Adaptation of an Instrument for Measuring School Coexistence in Middle School Students in Mexico

Cristina Vanessa Hernández De la Toba *
Joaquín Caso Niebla

[1] Instituto de Investigación y Desarrollo Educativo
de la Universidad Autónoma de Baja California

El presente estudio tuvo como propósito adaptar a la población mexicana el Cuestionario de Convivencia Escolar (CCE) elaborado por Díaz-Aguado, Martínez Arias y Martín (2010) dirigido a estudiantes del nivel secundaria, mediante el uso de procedimientos para la adaptación de instrumentos de medida (Díaz, 2015; Muñiz, Elosua y Hambleton, 2013); para la optimización de constructos complejos (Jornet, González-Such y Perales, 2012) y para la comprobación de invarianza de medición (Byrne, 2009; Cheung y Rensvold, 1999; Raju, Laffitte y Byrne, 2002). Se describen los resultados de cada una de las fases propuestas, producto de la aplicación de los CCE k=74 y CCE k=39 en muestras de Baja California N=660 y Querétaro N=419. El método se constituye de las fases: (a) Analítico-racional y (b) Empírica. A partir de la primera fase, se obtuvo el CCE adaptado k=74 con propiedades psicométricas aceptables (0.850 a 0,967); y como resultado de la segunda fase, se logró reducir el CCE a k=39 respetando los criterios de consistencia interna referidos por Jornet et al. (2012) y se comprobó la invarianza de la medición para las escalas de Fomento a la Convivencia Escolar, Acoso Escolar y Conductas Disruptivas del CCE adaptado y reducido.

Palabras clave: Adaptación, Optimización, Invarianza de medición, Validez, Constructo.

The present study was aimed to adapt the School Coexistence Questionnaire developed by Díaz-Aguado, Martínez Arias and Martín (2010) for middle school´s students using procedures for adapting measuring instruments (Díaz, 2015; Muñiz, Elosua, & Hambleton, 2013); for optimizing complex constructs (Jornet, González-Such, & Perales, 2012) and to test measurement invariance (Byrne, 2009; Cheung & Rensvold, 1999; Raju, Laffitte, & Byrne, 2002). Describes the results obtained in each of the phases proposed in the method, resulting form the application of the CCE k=74 and CCE k=39 in samples of Baja California N=660 and Querétaro N=419. The method is made up of the following two phases: (1) Analitic-rational (2) Empiric. At the first phase, it was obtained the adapted CCE k=74 with acceptables psychometric properties (0.850 a 0.967); at the second phase, it was reduced the CCE to K=39 accordance with the criteria of internal consistency indices referred by Jornet et al. (2012), and it was tested measurement invariance of the School Coexistence´s Promotion, Bullying and Disruptive Behavior Scales of adapted and reduced CCE.

Keywords: Adapting, Optimizing, Measurement invariance, Validity, Construct.

*Contacto: cristina.vanessa.hernandez.delatoba@uabc.edu.mx
issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 4 de julio de 2016
1ª Evaluación: 14 de septiembre de 2016
Aceptado: 1 de noviembre de 2016

1. Introducción

La convivencia escolar es un tema de interés internacional que está presente en diversos documentos normativos de políticas públicas (Jiménez, 2000). En el caso de México, desde 1982 a la fecha se han expedido acuerdos, leyes y programas ex profesos para el fomento de la convivencia en las escuelas. En los Acuerdos Secretariales 96, 97 y 98, se reglamentó la disciplina escolar para las escuelas primarias, secundarias técnicas y secundarias generales (Secretaría de Educación Pública, 1982). En 1993, se expidió la Ley General de Educación, en la que se cita en su artículo octavo, fracción tercera, que la educación contribuirá a la mejora de la convivencia humana (Secretaría de Educación Pública, 1993); en el 2007 se creó el Programa Nacional Escuela Segura con el propósito de consolidar a las escuelas públicas de educación básica como espacios seguros y confiables a través de la participación social y la formación ciudadana de los alumnos (Secretaría de Educación Pública, 2007). En el 2011 se llevó a cabo la reforma educativa a la educación básica en la que se incluyó el campo formativo de Desarrollo Personal y la Convivencia cuya finalidad es que los alumnos desarrollen el juicio crítico en favor de la democracia, la libertad, la paz, el respeto a las personas, a la legalidad y a los derechos (Secretaría de Educación Pública, 2011). En el 2011 y 2013 se realizaron importantes reformas al artículo tercero Constitucional asociadas con la convivencia en las escuelas (Cámara de Diputados del H. Congreso de la Unión, 2011; 2013). En el mismo 2013 se formuló el Programa Sectorial de Educación 2013-2018 (Secretaría de Educación Pública, 2013b), en el que se destacan líneas de acción asociadas al tema de la convivencia en las escuelas, tales como: la educación integral, la inclusión y la democracia; y por último, en el mismo 2013 la Secretaría de Educación Pública (SEP) puso énfasis en la construcción de ambientes donde los integrantes de la comunidad escolar convivan de manera pacífica, democrática e inclusiva (Secretaría de Educación Pública, 2013a).

Es innegable que el interés por mejorar la convivencia escolar en México es una preocupación sentida por las autoridades político-educativas. Sin embargo, un proceso de mejora implica el uso de la evaluación (Secretaría de Educación Pública, 2013b), puesto que la evaluación constituye una base de información para sensibilizar a la población sobre las condiciones en las que se presenta un fenómeno determinado o en las situaciones en las que se requiere poner atención (Rossi y Freeman, 1993), brinda orientación a los gobiernos para que puedan dirigir y ajustar estratégicamente sus políticas y programas hacia el logro de sus objetivos y metas (Cardozo, 2009; OCDE, 2011; Ponce, 2009).

No obstante, la evaluación de la convivencia es aún rudimentaria por ser un campo del conocimiento en construcción (Fierro et al, 2013) por ser un término utilizado por los actores del ámbito educativo no solamente para referirse a una gran diversidad de cuestiones; sino también, por ser estudiado desde perspectivas distintas (Fierro y Caso, 2013), por encontrarse en colindancia con otros referentes teóricos (Chaparro, Caso, Díaz y Urías, 2012) lo cual lo coloca en un constructo poliédrico, de múltiples caras (Ortega-Ruiz, Del Rey y Casas, 2013). Al respecto, en México se han documentado pocas experiencias de evaluación sobre la convivencia escolar a gran escala (Chaparro, Caso, y Fierro, 2012; Instituto Nacional para la Evaluación de la Educación, 2007, 2015; Ochoa y Salinas de la Vega, 2013; Valadez, 2008).

En este sentido, se reconoce la necesidad de contar con instrumentos de medición que ayuden a caracterizar este fenómeno de forma objetiva (Caso, Díaz y Chaparro, 2013), que permitan hacer distintas aproximaciones desde varias posturas teóricas y desde la perspectiva de diferentes actores (Fierro et al, 2013) y que generen conocimiento relacionado con los procesos implicados en el desarrollo de prácticas de convivencia inclusiva, democrática y no violenta que oriente la evaluación y el autodiagnóstico de escuelas, y que su vez motive el diseño de programas y propuestas de intervención (Fierro, 2011).

Por lo anterior, el objetivo central del presente estudio se basa en adaptar al contexto mexicano el Cuestionario de Convivencia Escolar para estudiantes de secundaria, el cual fue elaborado por el grupo de investigadores españoles integrado por Díaz Aguado, Martínez Arias y Martín (2010) con la mira de ser utilizado en estudios a gran escala, ya que estas evaluaciones tienen un peso cada vez mayor en temas de política nacional (Martínez, 2013). Sin embargo, al caracterizarse tales evaluaciones por recoger una mayor información, se pretende que sean utilizados instrumentos breves (Caso et al., 2013) a fin de simplificar el tiempo, el número de sesiones de la aplicación, el número de no respuestas (López-González, Tourón y Tejedor, 2012), y los efectos adversos como la fatiga y la falta de motivación derivados de un tiempo de administración excesivamente largo (Balluerka y Gorostiaga, 2012).

Bajo estas consideraciones, los objetivos específicos del presente estudio son: (a) Aplicar una metodología para la adaptación de instrumentos al Cuestionario de Convivencia Escolar; (b) Aplicar un conjunto de procedimientos para la optimización de la medida a la adaptación del Cuestionario de Convivencia Escolar; (c) Aportar evidencia de equivalencia métrica del Cuestionario de Convivencia Escolar que permita la comparación entre distintas muestras de México.

2. Método

De acuerdo con los objetivos del estudio, el método quedó conformado por las siguientes fases y etapas: 1) *fase analítico-racional*, misma que integra las etapas de adecuación del contenido, traducción y adecuación del formato, aplicación piloto y análisis psicométrico; 2) *fase empírica*, la cual está compuesta por las etapas de optimización, validación y de confirmación de la invarianza factorial (ver figura 1).

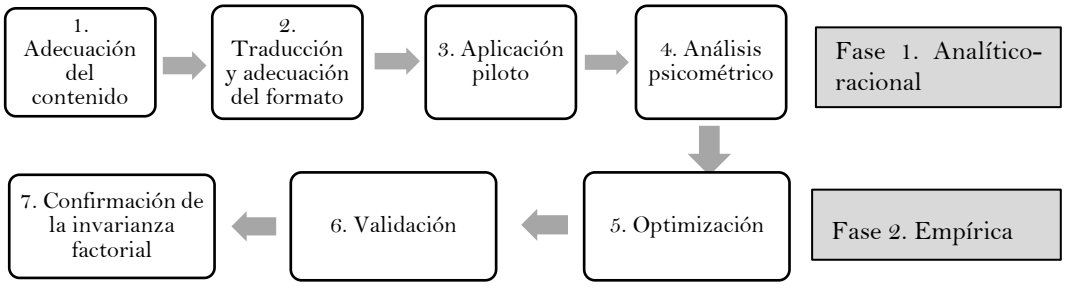


Figura 1. Fases y etapas del método del estudio
Fuente: Elaboración propia.

Las fases y etapas del método se describen a continuación.

2.1. Fase 1. Analítico-racional

La adaptación es un proceso mediante el cual es posible trasformar en -otro- un instrumento existente para emplearse en una población particular y ser utilizado de forma equivalente en una población distinta (APA, 2013), bajo el supuesto que el constructo medido por el test orientará hacia las mismas o similares predicciones e interpretaciones en la nueva cultura (Geisinger, 1994; Tanzer y Sim, 1999) y que permitirá realizar mediciones válidas en cada contexto o cultura (Klerk, 2008).

La adaptación es un proceso complejo que se halla expuesto a una gran cantidad de fuentes de error (Hambleton, 1996, 2005; Muñiz y Hambleton, 1996; Sireci, Patsula y Hambleton, 2005; Van de Vijver y Hambleton, 1996; Van de Vijver y Poortinga, 2005) que pueden ser controladas mediante la observancia de un marco de referencia (Cachón, 2007). En este sentido, y con la finalidad de evitar dichas fuentes de error, en esta fase se conjugaron una serie de procedimientos y directrices asociados con procesos de adaptación (Díaz, 2015; Hambleton, 1994, 1996; Muñiz y Hambleton, 1996; Muñiz et al., 2013).

2.1.1. Etapa 1. Adecuación del contenido

La presente fase tuvo como objetivo eliminar a aquellos ítems del CCE original *k=244* (Díaz-Aguado et al., 2010) que no cumplieran de manera satisfactoria con los criterios de pertinencia, congruencia y relevancia; así como determinar la suficiencia de los ítems para cada una de las escalas propuestas. El criterio de pertinencia hace referencia a la correspondencia conceptual entre los títulos de las dimensiones y los ítems que la integran. La congruencia es concebida como el grado en que el ítem es consecuente con la dimensión correspondiente, mientras que la relevancia se entiende como el grado en que la información contenida en el ítem es significativa para evaluar la dimensión a la que se adscribe; el criterio de suficiencia alude al grado en que se incluyen los ítems necesarios para evaluar cada dimensión.

- ✓ **Participantes.** Cuatro especialistas en temas y contenidos relacionados con la convivencia escolar, la psicometría y el diseño, desarrollo, validación y adaptación de instrumentos de medición psicológica y educativa.
- ✓ **Materiales**. La versión original del CCE para estudiantes (Díaz-Aguado et al., 2010), organizado en cinco dimensiones (ver tabla 1).

Tabla 1. Dimensiones y número de ítems del instrumento CCE original

DIMENSIÓN	K
Calidad de la convivencia escolar	42
Obstáculos a la convivencia en las relaciones entre estudiantes	35
Otros obstáculos a la convivência	62
Condiciones del centro educativo para construir y mejorar la convivencia	37
Otros indicadores	68
Total	*244*

Fuente: Elaboración propia.

- ✓ **Procedimiento.** Bajo el juicio de expertos a través del método de consenso (Escobar-Pérez y Cuervo-Martínez, 2008), las decisiones respecto a los criterios de pertinencia, congruencia, relevancia y suficiencia de los ítems de cada escala, fueron tomadas con base en el criterio referido por Herrera Rojas (1993), que existiera acuerdo entre más del 60% de los participantes.

2.1.2. Etapa 2. Traducción y adecuación del formato

Esta etapa tuvo como propósito transformar las palabras y expresiones de los ítems, opciones de respuesta e instrucciones que conforman el instrumento al contexto mexicano; así como modificar el formato en línea a formato de lápiz y papel debido a la limitación de los recursos informáticos. Las tareas realizadas fueron: (a) identificar los ítems expresados en forma de pregunta para ser transformados en afirmaciones; (b) asegurar la congruencia entre la base del ítem y sus opciones de respuesta; (c) valorar la estructura gramatical de los ítems originales de acuerdo al contexto de México y corregirlos; y (d) verificar la semántica de los ítems modificados. Respecto a la estructura gramatical, se valoró que el ítem se encontrara libre de errores gramaticales y que la sintaxis del enunciado fuera simple y común para la población destino. En relación a la semántica, se evalúo que las ideas y los significados transferidos al ítem traducido fueran iguales a los del ítem en el idioma fuente.

- ✓ **Participantes.** Se continuó con la colaboración del grupo de trabajo conformado en la etapa anterior considerando el conocimiento que poseían del instrumento en cuestión y por las limitaciones temporales que imperaban en el desarrollo del proyecto para conformar un nuevo comité. Aunque algunos autores señalan la conveniencia de expertos en lingüista para llevar a cabo las tareas indicadas en esta etapa (Costa y Brito, 2002; Díaz, 2015; Gaite, Ramírez, Herrera y Vázquez-Barquero, 1997; Muñiz et al., 2013; Ramada, Serra y Delclós, 2013; Villalobos, Arévalo y Rojas, 2012), el propósito de la misma no se vio afectado.
- ✓ **Materiales.** Se utilizó la versión del instrumento generada en la etapa anterior.
- ✓ **Procedimiento:** Se continuó con el juicio del grupo de expertos a través del método de consenso (Escobar-Pérez y Cuervo-Martínez, 2008). En todos los casos, las decisiones se tomaron al existir acuerdo entre más del 60% de los participantes (Herrera Rojas, 1993).

2.1.3. Etapa 3. Aplicación

El objetivo de esta etapa fue valorar el funcionamiento inicial de las escalas, ítems, instrucciones y opciones de respuesta.

- ✓ **Participantes.** Una muestra de 660 estudiantes del tercer grado de secundaria de los Municipios de Ensenada, Mexicali y Tijuana del Estado de Baja California, seleccionados de manera no probabilística (ver tabla 2).
- ✓ **Materiales.** Debido a que el Cuestionario de Convivencia Escolar formó parte de una estrategia evaluativa más amplia a cargo de la Unidad de Evaluación Educativa del Instituto de Investigación y Desarrollo Educativo de la UABC, denominada "Estrategia Evaluativa 2015, Factores asociados al aprendizaje" (Rodríguez, en prensa) se utilizaron los siguientes materiales: (a) Cuadernillo que contiene la batería de instrumentos incluida la nueva versión del

Cuestionario de Convivencia Escolar k=74; (b) Hojas de respuesta; (c) Hoja de incidencias.

- ✓ **Procedimiento.** Se diseñó y utilizó un procedimiento para cuidar la estandarización de la aplicación.

Tabla 2. Participantes de Baja California

		N	%
SEXO	Masculino	314	47,6
	Femenino	346	52,4
MUNICIPIO	Ensenada	105	15,9
	Tijuana	320	48,5
	Mexicali	235	35,6
MODALIDAD	General	234	35,5
	Privada	321	48,6
	Técnica	27	4,1
	Telesecundaria	78	11,8

Fuente: Elaboración propia.

2.1.4. Etapa 4. Análisis psicométrico (de los coeficientes de confiabilidad)

El objetivo de esta etapa fue obtener evidencias de confiabilidad del instrumento en cuestión.

- ✓ **Participantes.** Misma muestra de estudiantes de la etapa anterior.
- ✓ **Materiales**. (a) Hojas de respuestas; (b) Hoja de cálculo del programa Microsoft Excel. (c) Programas de análisis estadístico *Statistical Package for the Social Sciences* (SPSS) versión 19 y Programa R: *Project for Statistical Computing* versión 3.1.1.
- ✓ **Procedimiento.** Mediante el uso de los programas SPSS y R se realizaron los análisis de confiabilidad (alpha de Cronbach) y de discriminación (correlación punto biserial).

2.2. Fase 2. Empírica

2.2.1. Etapa 5. Optimización de la medida

El objetivo de esta etapa fue reducir la extensión del instrumento conservando los ítems clave que aportan mayor información respecto al constructo medido, aplicando el procedimiento propuesto por Jornet et al. (2012). Este procedimiento consiste en conservar sólo a los ítems clave portadores de información suficiente que permiten medir un determinado constructo a través de una escala con un elevado grado de confiabilidad y validez, equivalente a la que se obtendría con la aplicación total del cuestionario o de la escala original. Aunque tal procedimiento consta de seis pasos, sólo fueron aplicados los primeros cuatro por considerar que de continuar con la reducción del instrumento podría afectarse el entramado conceptual. Los pasos aplicados fueron: 1) Exploración del comportamiento de los reactivos de la escala original; 2) Análisis de valores perdidos y decisiones sobre imputación; 3) Selección de reactivos a partir de su contribución a la confiabilidad de la escala; y 4) Estimación del puntaje total de la versión reducida.

2.2.2. Etapa 6. Validación

El objetivo de esta etapa fue obtener evidencias de validez de constructo del micro-instrumento (versión reducida del instrumento), ya que es mediante la validez que se prueban las hipótesis acerca de relaciones teóricas (Messick, 1980) y se obtienen las evidencias que soportan las inferencias basadas en los puntajes obtenidos en el test (Contreras, 2000).

- ✓ **Participantes**. Misma muestra de la etapa tres.
- ✓ **Procedimiento**. Haciendo uso de la base de datos del programa SPSS de la muestra de Baja California, se contrastaron los modelos de medida construidos por el equipo de investigadores españoles aplicando la técnica de análisis factorial confirmatorio (AFC) por ser una técnica utilizada por excelencia para la validación de constructos (Pérez-Gil, Chacón y Moreno, 2000).

2.2.3. Etapa 8. Determinación de la invarianza factorial

El objetivo de esta etapa fue asegurar la invarianza o equivalencia factorial del CCE adaptado y reducido, a través de dos muestras -Baja California y Querétaro-. Este procedimiento, invarianza o equivalencia factorial, fue elegido por ser el indicado cuando las muestras pertenecen a diversas poblaciones (Gunnell, Wilson, Zumbo, Mack y Crocker, 2012).

- ✓ **Materiales.** (a) CCE adaptado y reducido; (b) hoja de respuestas, (c) procedimiento para la aplicación del cuestionario adaptado del procedimiento utilizado en la Estrategia Evaluativa 2015 de la Unidad de Evaluación Educativa (UEE-UABC) a fin de garantizar la estandarización de la aplicación; (d) reporte de incidencias; y (e) base de datos de la muestra de Baja California.
- ✓ **Participantes.** Misma muestra utilizada en la etapa tres, y una muestra de estudiantes del tercer grado del Estado de Querétaro N=419, (ver tabla 2).

Tabla 3. Participantes de Querétaro

		N	%
SEXO	Masculino	203	48,4
	Femenino	216	51,6
MUNICIPIO	San Joaquín	144	34,4
	Jalpán de Serra	77	18,4
	Vizcarrón Cadereyta de Montes	96	22,9
	Ahuacatlán de Guadalupe	102	24,3

Fuente: Elaboración propia.

- ✓ **Procedimiento**. Se utilizó la técnica de análisis factorial confirmatorio multigrupo (AFCMG) ya que facilita la investigación del grado en que las medidas son invariantes entre los grupos (Chen, 2008). En concreto, esta técnica permite verificar si los niveles de las variables observadas de la variable latente entre los grupos tienen los mismos puntajes esperados en la medida (Drasgow y Kanfer, 1985).

  La prueba de la invarianza de la medida consistió en la comparación de los índices de ajuste los modelos: base *(configural)*, débil *(weak o metric)*, fuerte *(strong o scalar)*, y estricto *(strict)* aplicando restricciones de igualdad (Byrne, 2009; Cheung y Rensvold, 1999; Raju, Laffitte y Byrne, 2002) para cada una de las

escalas del CCE adaptado. Estos índices de ajuste se obtuvieron aplicando el AFCMG en cada uno de los modelos con apego a las siguientes restricciones sugeridas por (Hirschfeld y Von Brachel, 2014):

- El modelo base o *configural* se mantuvo libre de restricciones. En este modelo, se supone que el patrón de cargas es similar en todos los grupos, pero las magnitudes de los parámetros (cargas, interceptos,varianzas) pueden variar;
- En el modelo débil (*weak o metric)* se restringieron las cargas factoriales a ser iguales;
- En el modelo fuerte (*strong o scalar)* se restringieron las cargas factoriales y los interceptos;
- Dado que el modelo estricto (*strict)* exige que las cargas factoriales, los interceptos y las varianzas residuales deben ser restringidas y que la invarianza estricta requiere del uso de la theta-parametrización para identificar los parámetros del modelo (Hirschfeld y Von Brachel, 2014), éste no se configuró debido a que la identificación de los parámetros que ofrece la versión del progama R utilizado, era distinta a la sugerida por el autor.

Al realizar el AFCMG en cada uno de los modelos de las escalas se obtuvieron los índices de ajuste absolutos y comparativos RMSEA (≤ 0,06), CFI (≥ 0,95) y TLI (≥ 0,95) (Abad, Olea, Ponsoda, y García, 2011; Hu y Bentler, 1999).

Una vez confirmados los índices de ajuste de los modelos, se obtuvieron nuevamente los índices de ajuste de los modelos anidados (base, débil y fuerte), permitiendo la contrastación de los índices de ajuste (CFI y RMSEA) y la determinación de la invarianza mediante la diferencia entre los índices de ajuste ≤ 0,010 (Hirschfeld y Von Brachel, 2014).

3. Resultados

A continuación se presentan los resultados obtenidos en cada una de las fases y etapas correspondientes del presente estudio.

3.1. Fase 1. Analítico-racional

3.1.1. Etapa 1. Adecuación del contenido

Al realizar el análisis de las dimensiones del CCE original mediante el criterio de pertinencia, se observó que al interior de las dimensiones se aglomeraban distintos constructos. Por lo que, bajo la definición de convivencia escolar como un conjunto de relaciones entre los actores de una comunidad educativa -alumnos, profesores, directivos y padres de familia- (Bazdresch, 2009; Díaz-Aguado, s.f.; Fernández, 1998; Fierro et al., 2013; Tuvilla, s.f.) que implican la comprensión de las diferencias, el aprecio a la interdependencia y la pluralidad, el aprendizaje para enfrentar los conflictos de una manera positiva y la promoción del entendimiento mutuo y la paz mediante la participación democrática (Carbajal, 2013), se optó por conservar de las dimensiones originales del CCE sólo cuatro rasgos para su operacionalización agrupados en dimensiones y subdimensiones (ver tabla 4). Con base en tal decisión, el comité juzgó

que la mayoría de los ítems no formaban parte del constructo - los cuatros rasgos - a medir y que además, otra buena parte aunque sí correspondían a la dimensión, había ítems suficientes que permitían operacionalizarlo. Por lo tanto, de los 244 ítems que conformaban el CCE original, se eliminaron 172, 121 por no cumplir con el criterio relevancia y 51 por no apegarse al criterio de congruencia, conservándose 72 de los ítems originales (ver tabla 5). En el caso de las dimensiones de Clima Escolar y Acoso Escolar se consideró la carencia de elementos del constructo indispensables. Así que, de acuerdo con el criterio de suficiencia, se determinó la inclusión de dos ítems (ver tabla 4).

Tabla 4. Escalas del instrumento original y del instrumento modificado

ORIGINAL		**ADAPTADA**		
K	**Escala/dimensión**	**K**	**Escala/dimensión**	**Subdimensión**
42	Calidad de la convivencia escolar	**13 (+1)**	Clima escolar	Relación entre profesores y estudiantes Relación entre la familia y la escuela
35	Obstáculos a la convivencia en las relaciones entre estudiantes	**31 (+1)**	Acoso escolar	Victimización de acoso escolar Participación en acoso escolar
62	Otros obstáculos a la convivencia	**16**	Conductas disruptivas	Conductas disruptivas en el aula Conductas agresivas entre profesores y estudiantes
37	Condiciones del centro educativo para construir y mejorar la convivencia	**14**	Fomento a la convivencia escolar	Respeto a la diversidad Enseñanza de preceptos de convivencia escolar
68	Otros indicadores	Se eliminó		

Nota: +1= inclusión de nuevo reactivo.

Fuente: Elaboración propia.

Si bien, la reducción del instrumento original fue considerable, no se descarta que se trata de un proceso de adaptación bajo la categoría de ensamblaje. Según Tanzer y Sim (1999) y Van de Vijver y Hambleton (1996), el ensamblaje es un enfoque de adaptación que equivale a compilar un nuevo instrumento por haber sido modificado a -profundidad- y en el que se resalta la pertinencia cultural del instrumento, al considerar aspectos del constructo que son importantes para la cultura objetivo pero no para el instrumento original.

Tabla 5. Cantidad de ítems eliminados por no reunir los criterios de relevancia y congruencia

DIMENSIÓN	**ÍTEMS ELIMINADOS**			
	k original	**No relevancia**	**No congruencia**	**Total**
Calidad de la convivencia escolar/Clima Escolar	42	28	2	30
Obstáculos a la convivencia en las relaciones entre estudiantes/Acoso Escolar	35	3	2	5
Otros obstáculos a la convivencia/Conductas Disruptivas	62	34	12	46
Condiciones del centro educativo para construir y mejorar la convivencia/Fomento a la convivencia escolar	37	16	7	23
Otros indicadores	68	40	28	68
Total	*244*	*121*	*51*	*172*

Fuente: Elaboración propia.

3.1.2. Etapa 2. Traducción y adecuación de formato

El comité observó que algunos de los ítems estaban expresados en forma de pregunta, lo que derivó en la modificación del ítem-pregunta a ítem-afirmación. También se identificó que todos los ítems eran objeto de modificación en cuestiones de gramática y semántica de acuerdo con el idioma castellano de México (ver tabla 6). En la tabla 7 se presentan algunos ejemplos de los cambios señalados.

Tabla 6. Número de ítems modificados

DIMENSIÓN	ÍTEMS MODIFICADOS			
	Por expresarse en pregunta	Por aspectos gramaticales	Por aspectos semánticos	Total
Clima escolar	0	12	2	12
Acoso escolar	8	22	3	30
Conductas disruptivas	0	20	0	20
Fomento a la convivencia	0	10	2	10
Total	*8*	*64*	*7*	*72*

Fuente: Elaboración propia.

Tabla 7. Ítems modificados por expresarse en pregunta, por aspectos gramaticales y semánticos

ÍTEM ORIGINAL	EXPRESADO EN PREGUNTA	ASPECTO GRAMATICAL	ASPECTO SEMÁNTICO	ÍTEM MODIFICADO
Los estudiantes se ayudan entre sí, aunque no sean amigos		1		Los estudiantes de esta escuela se ayudan entre sí, aunque no sean amigos(as)
Hay peleas entre estudiantes		1		En esta escuela es común que se peleen los estudiantes
Los estudiantes que pertenecen a distintos grupos o pandillas se llevan bien		1	1	Los estudiantes de esta escuela se llevan bien sin importar que pertenezcan a diferentes grupos
Hay bandas en el centro que perjudican la convivencia		1	1	En esta escuela hay estudiantes que afectan la convivencia.
Tiene manía a algunos estudiantes		1	1	A los profesores les caen mal algunos estudiantes
¿Has recibido mensajes a través de Internet o de teléfono móvil en los que te insultaran, amenazaran, ofendieran o asustaran?	Sí			Mis compañeros(as) me envían mensajes a través de internet o teléfono celular en los que me insultan u ofenden
¿Han difundido fotos o imágenes tuyas por Internet o teléfono móvil para Obligarte a hacer después algo que no querías con amenazas?	Sí			Mis compañeros(as) han difundido videos o imágenes mías por Internet o teléfono celular para utilizarlas en mi contra

Fuente: Elaboración propia.

En relación a la congruencia entre instrucciones-ítem-opciones de respuesta, debido a las modificaciones antes realizadas, fue necesario cambiar todas las opciones de respuesta e instrucciones del cuestionario.

3.1.3. Etapa 3 y 4. Aplicación y análisis de los coeficientes de confiabilidad del instrumento

La confiabilidad se ha utilizado para referirse a los coeficientes de confiabilidad de la Teoría Clásica de los Test (TCT) entendidos entre otros, como la consistencia interna del conjunto de ítems, si se analiza el grado en que los distintos ítems miden cierto constructo (American Education Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Cuando un test es traducido o adaptado para su uso en un nuevo grupo lingüístico o cultural, los índices de confiabilidad deben calcularse y estudiarse (American Education Research Association et al., 2014; Geisinger, 1994).

A partir de la aplicación piloto, se calcularon los coeficientes de consistencia interna (alpha de Cronbach) para cada una de las escalas, obteniendo puntuaciones que van de 0.850 a 0.967, y se obtuvieron los índices de correlación $r_{bis} \geq 0.20$ (ver tabla 8). A partir de los criterios de Contreras (2000), $r_{bis} \geq 0.20$ y George y Mallery (2003), alfa de Cronbach >.9 excelente y >.8 bueno, los valores resultantes se consideraron como aceptables.

Tabla 8. Índices de consistencia interna y de discriminación de las escalas

ESCALAS/DIMENSIONES	K	ALPHA DE CRONBACH	$R_{BIS} \leq 0.20$
Clima escolar	13	0,850	0
Acoso Escolar	31	0,967	0
Conductas disruptivas	20	0,932	0
Fomento a la convivencia escolar	10	0,884	0

Fuente: Elaboración propia.

3.2. Fase 2. Reductiva

3.2.1. Etapa 5. Optimización de la medida

Paso 1. Se identificaron los ítems que presentaban escasa variabilidad eliminándose aquellos elementos con una distribución atípica. Como criterio de exclusión se estableció una variabilidad mayor a 80%, por lo que se eliminaron siete ítems, tres de la Escala de Clima Escolar y cuatro de la Escala de Conductas Disruptivas (ver tabla 9).

Paso 2. En esta fase se analizaron los valores perdidos con la perspectiva de que su existencia fuera de carácter aleatorio y no a un patrón de no respuestas sistemático. En las variables o ítems donde se registraron valores perdidos menores al 30%, fueron sustituidos por la mediana, tal como lo sugirieron Jornet et al. (2012). Cabe mencionar, que en ninguno de los casos se observó que los valores perdidos excedieran el límite permitido.

Paso 3. En esta fase se buscó la reducción de los ítems a partir de la identificación de los ítems redundantes. Para poder identificarlos, se buscó a aquellos ítems con una correlación mayor a 0,50. Atendiendo estos criterios se eliminaron 28 ítems por presentar una correlación ≥ 0,50. A pesar de haber identificado otros ítems que se apegaban al criterio se determinó conservarlos debido a que al hacerlo se afectaba la definición del constructo. El presente análisis dejó como saldo la eliminación de dos ítems de la Escala de Clima Escolar, 16 de la Escala de Acoso Escolar, seis de la escala

de Conductas Disruptivas y cuatro de la Escala de Fomento a la Convivencia Escolar (ver tabla 9).

Tabla 9. Optimización de la medida de las escalas

ESCALA CLIMA ESCOLAR					
Versión	**K**	**alfa**	**rbis**	**ítems eliminados**	**Criterios de exclusión**
1	13	0,8500	1	---	---
2	10	0,7800	0,970	3	variabilidad > 80%
3	8	0.,8200	0,938	2	r≥ .5 y análisis conceptual
ESCALA ACOSO ESCOLAR					
Versión	**K**	**alfa**	**rbis**	**ítems eliminados**	**Criterios de exclusión**
1	31	0,967421	1	---	---
2	15	0,9502573	0,970	16	r≥ .5 y análisis conceptual
ESCALA CONDUCTAS DISRUPTIVAS					
Versión	**K**	**alfa**	**rbis**	**ítems eliminados**	**Criterios de exclusión**
1	20	0,9317074	1	---	---
2	16	0,9225304	0,983	4	variabilidad > 80%
3	10	0,8948949	0,967	6	r ≥ .5 análisis conceptual
ESCALA FOMENTO A LA CONVIVENCIA ESCOLAR					
Versión	**K**	**alfa**	**rbis**	**ítems eliminados**	**Criterios de exclusión**
1	10	0,8837126	1	---	---
2	6	0,713786	0,896	4	r ≥ .5 análisis conceptual

Nota: --- No aplica: r= correlación; rbis=correlación punto biserial.
Fuente: Elaboración propia.

Fase 4 y 5. Se estimó la puntuación total del cuestionario mediante la suma total de las puntuaciones de los ítems y se exploró la calidad de la versión dos a través del índice de consistencia interna en cada una de las escalas a fin de medir el tamaño de la pérdida en confiabilidad de la versión obtenida de la fase anterior, cuidando que ésta no excediera el 10% del alfa inicial - pérdida del alfa ≤ 10% - (ver tabla 9).

Fase 6. En la metodología propuesta por Jornet et al. (2012) se sugiere esta segunda etapa de reducción. Sin embargo, se determinó obviarla debido a que se consideró que en caso de continuar con la reducción del cuestionario se podría afectar su entramado conceptual.

3.2.2. Etapa 6. Validación

Se aplicó el análisis factorial confirmatorio (AFC) utilizando el método de estimación de mínimos cuadrados no ponderados (ULS) sobre la matiz de correlaciones policórica, por motivo de la naturaleza ordinal de las variables observadas (Luo, 2011). A partir del AFC y en consideración a lo expuesto por Kim y Bentler (2006), se obtuvieron los índices de ajuste absoluto SRMR (*Standarized Root Mean Square Residual*) y el RMSEA (*Root Mean Square error aproximation*), así como los índices de ajuste comparativo CFI (*Comparative Fit Index*) y TLI (*Índice de Tucker-Lewis*). El ajuste del modelo se determinó con base en los índices de bondad de ajuste absoluto SRMR (≤ 0,08) y RMSEA (≤ 0,06) y los índices de bondad de ajuste comparativo CFI (≥ 0,95) y TLI (≥ 0,95) (Abad, et al., 2011; Hu y Bentler, 1999).

Como se observa en la tabla 10, los índices para cada una de las escalas: Clima Escolar, Acoso Escolar, Conductas Disruptivas y Fomento a la Convivencia Escolar reflejan un buen ajuste en cada uno de los modelos ⁅CFI (≥ 0,95) y TLI (≥ 0,95); SRMR (≤0,08); y RMSEA ≤ 0,07⁆; aunque en las escalas de Acoso Escolar y Fomento a la Convivencia Escolar el RMSEA no haya resultado significativo. Al respecto, Kim y Bentler (2006)

sostienen que la valoración de ajuste de los datos puede realizarse mediante el uso de por lo menos dos tipos de índices diferentes (absoluto, incremental o de parsimonia), por ejemplo TLI y SRMR, RMSEA y SRMR o CFI y SRMR (Hu y Bentler, 1999).

Tabla 10. Índices de ajuste de las escalas del Cuestionario Convivencia Escolar adaptado y reducido

ÍNDICE DE AJUSTE	CLIMA ESCOLAR	ACOSO ESCOLAR	CONDUCTAS DISRUPTIVAS	FOMENTO A LA CONVIVENCIA ESCOLAR
Chi-cuadrado	172,487	180,500	127,292	2,951
Grados de libertad	19	89	34	4
CFI	0,950	0,997	0,987	1,000
TLI	0,926	0,996	0,983	1,001
SRMR	0,020	0,048	0,059	0,017
RMSEA	0,111	0,039	0,065	0,000
Intervalo de confianza	0,096-0,126	0,031-0,048	0,053-0,077	0,000-0,051
Valor p RMSEA	0,000	0,983	0,022	0,944

Fuente: Elaboración propia.

3.3. Fase 3. Equivalencia

3.3.1. Etapa 5. Determinación de la invarianza factorial

Con la finalidad de asegurar la invarianza o equivalencia factorial de las escalas entre las muestras de Baja California y Querétaro se generó una sola base de datos, misma que permitió el cálculo de los índices de ajuste absoluto, así como los índices de ajuste de los modelos anidados. Para determinar el ajuste de los modelos, se consideraron ajustes de tipo exacto o absoluto, (chi-cuadrada no significativa) (Hirschfeld y Von Brachel, 2014; Magnus, 2005) y ajustes de tipo aproximado o comparativo (CFI ≥ 0,95, TLI ≥ 0,95 y RMSEA ≤ 0,06) (Abad et al., 2011; Hu y Bentler, 1999). No obstante, debido a que la chi cuadrada es sensible al tamaño de la muestra (Byrne y Van de Vijver, 2010), ésta fue utilizada con precaución.

Para el caso de la Escala de Clima Escolar, se observa que los índices de ajuste absolutos para los modelos *configural* y *weak* fueron aceptables. Sin embargo, aunque en el modelo *strong* el RMSEA estuvo muy por encima de los valores esperados, se obtuvo un TLI ≥ 0,95 y un CFI ≥ 0,99 el cual puede interpretarse como un criterio confiable (Hu y Bentler, 1999). Respecto a la Escala Fomento a la Convivencia Escolar, todos los índices de ajuste absoluto de los modelos (*configural*, *weak* y *strong*) resultaron aceptables, con excepción de la chi-cuadrada (ver tabla 11).

En la escala de Acoso Escolar, en los tres modelos el CFI fue aceptable y el TLI en los modelos *weak* y *strong* estuvieron dentro del rango (≥ 0,95), con excepción del modelo *configural*. En el caso de RMSEA, el modelo *configural* y el *strong* se mantuvieron en el índice esperado (≤ 0,06), con la excepción del modelo *weak*. Sin embargo, a pesar de los contrastes entre los índices resultantes, se considera que los modelos tienen un buen ajuste, ya que el CFI en todos los casos fue aceptable (Hu y Bentler, 1999); y por último, en la escala de Conductas Disruptivas, los índices de ajuste en los tres modelos del CFI y TLI fueron aceptables (≥ 0,95), y no para el caso del RMSEA, ya que en todos los modelos se obtuvieron índices superiores a lo esperado (≤ 0,06), ver tabla 11.

Tabla 11. Índices de ajuste absoluto de los modelos de las escalas del Cuestionario

ESCALA	MODELOS	CHISQ	DF	P	CFI	TLI	RMSEA
Clima Escolar	Configural	55,972	22	0,000	0,993	0,982	0,054
	Weak	67,707	35	0,001	0,993	0,989	0,042
	Strong	82,799	48	0,001	0,993	0,993	0,992
Fomento a la Convivencia Escolar	Configural	18,001	8	0,021	0,998	0,995	0,048
	Weak	16,18	11	0,135	0,999	0,998	0,030
	Strong	23,298	19	0,224	0,999	0,999	0,020
Acoso Escolar	Configural	955,522	178	0,000	0,955	0,946	0,090
	Weak	903,669	191	0,000	0,958	0,954	0,954
	Strong	949,942	219	0,000	0,957	0,959	0,079
Conductas Disruptivas	Configural	354,059	68	0,000	0,991	0,988	0,088
	Weak	409,304	76	0,000	0,989	0,987	0,090
	Strong	392,844	94	0,000	0,990	0,991	0,077

Nota: chisq=chi cuadrada; df=grados de libertad; p=nivel de significancia; CFI=*Comparative Fit Index;* TLI=*Índice de Tucker-Lewis*; RMSEA=*Root Mean Square error aproximation.*
Fuente: Elaboración propia.

Una vez asegurados los índices de ajuste absoluto, se procedió al cálculo de los ajustes de los modelos anidados. En el caso de la Escala de Clima Escolar, para los modelos *weak* y *strong* se obtuvieron puntajes aceptables en todos los índices y una chi-cuadrada no significativa. Cabe mencionar que aunque en el modelo *configural*, el CFI resultó muy por debajo del puntaje esperado, el resto de los índices se mantuvieron en el rango esperado. Para el caso de la Escala de Fomento a la Convivencia, todos los índices resultantes se encuentran dentro de los criterios establecidos (ver tabla 12).

Respecto a la Escala de Acoso Escolar, se observa que en los tres modelos el CFI y TLI resultaron aceptables. En el caso del RMSEA sólo el modelo *strong* estuvo dentro del criterio (≤ 0.06); y en el caso de la Escala de Conductas Disruptivas, al igual que la escala anterior, los índices de ajuste de los modelos anidados resultaron aceptables para el CFI y TLI (≥ 0.95), mientras que el RMSEA se ubicó fuera de los parámetros establecidos para los tres casos restantes (ver tabla 12).

Tabla 12. Índices de ajuste de los modelos anidados de las escalas de Cuestionario

ESCALA	MODELOS	CHISQ	DF	P	CFI	TLI	RMSEA
Clima Escolar	Configural	24,620	22	0,316	0,000	0,999	0,015
	Weak	43,488	35	0,154	0,999	0,998	0,021
	Strong	52,057	48	0,319	0,999	0,999	0,013
Fomento a la Convivencia Escolar	Configural	6,439	8	0,598	1,000	1,001	0,000
	Weak	7,46	11	0,761	1,000	1,001	0,000
	Strong	11,844	19	0,892	1,000	1,001	0,000
Acoso Escolar	Configural	649,193	178	0,000	0,991	0,989	0,070
	Weak	772,487	191	0,000	0,989	0,988	0,075
	Strong	699,008	219	0,000	0,991	0,991	0,064
Conductas Disruptivas	Configural	354,059	68	0,000	0,991	0,988	0,088
	Weak	409,304	76	0,000	0,989	0,987	0,090
	Strong	392,844	94	0,000	0,990	0,991	0,077

Nota: chisq=chi cuadrada; df=grados de libertad; p=nivel de significancia; CFI=*Comparative Fit Index;* TLI=*Índice de Tucker-Lewis*; RMSEA=*Root Mean Square error aproximation.*
Fuente: Elaboración propia.

Una vez analizados y aceptados los índices de ajuste anidados según los rangos esperados, se procedió a calcular las diferencias entre dichos modelos buscando asegurar la evidencia de invarianza entre el modelo configural y el débil o métrico, ya que dicha

evidencia es fundamental para establecer la invarianza de medición y la interpretación conceptual de un instrumento (Pickering y Plotnikoff, 2009).

Con base lo anterior, en la Escala de Clima Escolar se observó que la diferencia del CFI entre el modelo *configural-weak*, no presentó índices de ajuste aceptables, mientras que la diferencia entre los modelos *weak-strong* resultó aceptable (CFI.delta ≤ 0,010). Para el caso de la Escala de Fomento a la Convivencia Escolar, la diferencia entre los modelos *configural-weak* y *weak-strong*, resultaron dentro de los parámetros establecidos (ver tabla 13).

Tabla 13. Comparación de modelos anidados de la Escala Clima Escolar

ESCALA	MODELOS	CHISQ. DELTA	DF. DELTA	P. DELTA	CFI. DELTA	TLI. DELTA	RMSEA. DELTA
Clima Escolar	Configural-Weak	18,868	13	-0,162	0,999	-0,001	0,006
	Weak-Strong	8,569	13	0,165	0,000	0,001	-0,008
Fomento a la Convivencia Escolar	Configural-Weak	1,021	3.000	0,163	0,000	0,000	0,000
	Weak-Strong	4,384	8.000	0,131	0,000	0,000	0,000
Acoso Escolar	Configural-Weak	123,294	13	0,000	-0,002	-0,001	0,005
	Weak-Strong	-73,479	28	0,000	0,002	0,003	-0,011
Conductas Disruptivas	Configural-Weak	55,245	8	0,000	-0,002	-0,001	0,002
	Weak-Strong	-16,46	18	0,000	0,001	0,004	-0,013

Nota: chisq.delta=diferencia entre la chi cuadrada; df.delta=diferencia entre los grados de libertad; p.delta=diferencia entre el nivel de significancia; CFI.delta=diferencia entre el *Comparative Fit Index;* TLI.delta= diferencia entre el *Índice de Tucker-Lewis*; RMSEA.delta= diferencia entre el *Root Mean Square error aproximation.*
Fuente: Elaboración propia.

Acerca de la Escala de Acoso Escolar, en la tabla 13 se observa que las diferencias entre los modelos *configural-weak y weak-strong* se encuentran dentro del rango establecido CFI.delta (≤ 0,010); y por último, en la Escala de Conductas Disruptivas, una vez realizada la comparación de las diferencias del CFI.delta entre los modelos *configural-weak* y *weak-strong*, se observa que ambas comparaciones permiten asegurar la equivalencia de la medida (CFI.delta ≤ 0,10), ver tabla 13.

4. Discusión

El objetivo central de este trabajo ha consistido en mostrar la aplicación de metodologías para la adaptación, optimización y validación del Cuestionario de Convivencia Escolar (Díaz-Aguado, et al., 2010). Con base en los principales resultados, es posible decir que mediante el procedimiento de adaptación aplicado, se logró obtener una versión del Cuestionario de Convivencia Escolar (adaptado) k=74, útil para caracterizar muestras de estudiantes en el contexto de Baja California, puesto que los índices de consistencia interna van desde un nivel bueno a excelente (de 0,850 a 0,967), según lo señalado George y Mallery (2003). De la misma manera, el procedimiento de optimización de la medida propuesto por Jornet, et al. (2012) resultó óptimo, ya que permitió reducir el Cuestionario a k=39 manteniendo los índices de consistencia interna dentro del rango esperado (≤ 10% al índice inicial).

Bajo estos resultados favorables, y con la finalidad de contar con un instrumento para ser aplicado no sólo en Baja California sino en nuestro país, el Cuestionario k=39 fue sometido a un procedimiento de invarianza de medición. Con base en los resultados

obtenidos en la fase de validación empírica, es posible asegurar que en el caso de las Escalas de Fomento a la Convivencia Escolar, Acoso Escolar y Conductas Disruptivas que integran el Cuestionario de Convivencia Escolar, pueden ser utilizadas en estudios a gran escala para describir de manera general la situación que se vive en las escuelas mexicanas respecto al tema de la convivencia. Tal aseveración, se sustenta en que las diferencias entre los modelos restringidos *configural-weak* se encuentran dentro del rango establecido CFI.delta (≤ 0,010), evidencia fundamental para establecer la invarianza de medición y la interpretación conceptual de un instrumento (Pickering y Plotnikoff, 2009). No obstante, se sugiere utilizar con mayor cautela los resultados que se obtengan a partir de la Escala de Clima Escolar, ya que aunque la diferencia entre los modelos *weak-strong* se mantuvo dentro del parámetro señalado, no fue el caso para los modelos *configural-weak*.

No obstante con todo lo alentador de los resultados obtenidos, queda aún pendiente por realizar la aplicación de ambos instrumentos k=74 y k=39 en muestras de estudiantes más amplias, puesto que resulta de gran relevancia el documentar sus propiedades psicométricas.

Referencias

Abad, F., Olea, J., Ponsoda, V. y García, C. (2011). *Medición en Ciencias Sociales y de la Salud*. Madrid: Síntesis.

American Education Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standars for educational and pshycological testing*. Washington, D.C.: American Education Research Association.

American Psychological Association. (2013). *Going international: A practical guide for psychologists. Using translated/adapted measurement scales internationally*. Nueva York, NY: Autor.

Baena, G. (1998). *Técnicas de investigación*. Ciudad de México: Editores Mexicanos Unidos.

Balluerka, N. y Gorostiaga, A. (2012). Elaboración de versiones reducidas de instrumentos de medida: Una perspectiva práctica. *Intervención Psicosocial* , *21*(1), 103-110. doi: 10.5093/in2012v21n1a7

Bazdresch, M. (2009). La vida cotidiana escolar en la formación valoral, un caso. *REICE. Revista sobre Calidad, Eficacia y Cambio en Educación* , *7*(2), 49-71.

Byrne, B. (2009). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Nueva York, NY: Taylor and Fancis Group.

Byrne, B. y Van de Vijver, F. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of non equivalence. *International Journal of Testing* , *10*(2), 107-132. doi: 10.1080/15305051003637306

Cámara de Diputados del H. Congreso de la Unión. (2011). *Diario Oficial de la Federación*. Ciudad de México: Diario Oficial de la Federación.

Cámara de Diputados del H. Congreso de la Unión. (2013). *Diario Oficial de la Federación*. Ciudad de México: Diario Oficial de la Federación.

Cachón, C. (2007). *La función de la equivalencia en el proceso de la medición intercultural. Memoria electrónica COMIE*. Ciudad de México: COMIE.

Carbajal, P. (2013). Convivencia democrática en las escuelas. Apuntes para una reconceptualización. *Revista Iberoamericana de Evaluación Educativa* , *6*(2), 13-35.

Cardozo, M. (2009). La institucionalización de una cultura de la evaluación en la administración pública mexicana: Avances y desafíos pendientes. *Convergencia. Revista de Ciencias Sociales*, *16*(49), 175-198.

Caso, J., Díaz, C. y Chaparro, A. (2013). Aplicación de un procedimiento para la optimización de la medida de la convivencia escolar. *Revista Iberoamerican de Evaluación Educativa*, *6*(2), 137-145.

Chaparro, A., Caso, J. y Fierro, C. (2012). *Validación psicométrica de indicadores de convivencia democrática, inclusiva y pacífica. Reporte de resultados.* Ciudad de México: CONCYTEG.

Chaparro, A., Caso, J., Díaz, C. y Urías, E. (2012). *Instrumentos para el diagnóstico e intervención en escuelas basados en indicadores de convivencia democrática, inclusiva y no violenta*. Ensenada: Universidad Autónoma de Baja California.

Chen, F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropiate comparisons in cross-cultural research. *Jorunal of Personality and Social Psychology* , *25*(1), 1005-1018. doi: 10.1037/a0013193

Cheung, G. y Rensvold, R. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *1*, 1-27. doi: 10.1177/014920639902500101

CONEVAL. (2012). *Informe de pobreza y evaluación en el estado de Sonora*. Ciudad de México: Autor. Recuperado de goo.gl/bvhAFW

CONEVAL. (2012). *Porcentaje de población en situación de probreza extrema según entidad federativa*. Ciudad de México: Autor. Recuperado de goo.gl/vJBXct

Contreras, L. A. (2000). *Desarrollo y pilotaje de un examen de español para la educación primaria* (Tesis de maestría). Universidad Autónoma de Baja California, México.

Costa, N. y Brito, E. (2002). Adaptación cultural de instrumentos utilizados en salud ocupacional. *Revista Panamericana de Salud Pública*, *11*(2), 109-111. doi:10.1590/S1020-49892002000200007

Delval, J. (1992). *Aprender a aprender. I. El desarrollo de la capacidad de pensar*. Madrid: Alhambra Longman.

Díaz Aguado, M. J., Martínez Arias, R. y Martín, J. (2010). *Estudio estatal de convivencia escolar de la educación secundaria 2007-2009*. Madrid: Ministerio de Educación.

Díaz, C. D. (2015). *Adaptación de un instrumento para la medición de la convivencia escolar* (Tesis de maestría). Universidad Autónoma de Baja California, México.

Drasgow, F. y Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*(4), 662-680.

Elosua, P., Mujika, J., Almeida, L. y Hermosilla, D. (2014). Procedimientos analítico-racionales en la adaptación de test. Adaptación al español de la batería de pruebas de razonamiento. *Revista Latinoamericana de Psicología* , *45*(2), 117-126. doi: 10.1016/S0120-0534(14)70015-9

Escobar-Pérez, J. y Cuervo-Martínez, A. (2008). Validez de contenido y juicio de expertos: Una aproximación a su utilización. *Avances en Medición*, *6*, 27-36.

Fernández, I. (1998). *Prevención de la violencia y resolución de conflictos: Clima escolar con factor de calidad*. Madrid: Narcea.

Fierro, C. (2011). *Convivencia democrática e inclusiva. Una perspectiva para gestionar la seguridad escolar*. Recuperado de http://basica.sep.gob.mx/escuelasegura/pdf/congresoBuenasPrac/convivencia.pdf

Fierro, C. y Caso, J. (2013). Presentación del monográfico. *Revista Iberoamericana de Evaluación Educativa* , *6*(2), 7-12.

Fierro, C., Tapia, G., Martínez-Parente, R., Macouzet, M. y Jiménez, M. (2013). Conversando sobre la convivencia en la escuela: Una guía para el autodiagnóstico de la convivencia escolar. *Revista Iberoameericana de Evaluación Educativa*, *2*(6), 103-124.

Fierro, C., Tapia, G., Fortoul, B., Martínez-Parente, R., Macouzet, M. y Jiménez, M. (2013). Conversando sobre la convivencia en la escuela: Una guía para el auto-diagnóstico de la convivencia escolar. *Revista Iberoamericana de Evaluación Educativa*, *2*(6), 103-124.

Gaite, L., Ramírez, N., Herrera, S. y Vázquez-Barquero, J. (1997). Traducción y adaptación transcultural de instrumentos en psquiatría: Aspectos metodológicos. *Archivos de Neurobiología*, *10*(2), 189-214.

Geisinger, K. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304-312. doi: 10.1037/1040-3590.6.4.304

George, D. y Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. Boston, MA: Allen & Bacon.

Gunnell, K., Wilson, P., Zumbo, B., Mack, D. y Crocker, P. (2012). Assesing psychological need satisfaction in exercise contexts: Issues of score invariance, item modification, and context. *Measurement in Physical Education and Exercise Science*, *16*(3), 219-236. doi: 10.1080/1091367X.2012.693340

Hambleton, R. K. (1996). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*, 229–240.

Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting test into multiple languages and cultures. En R. Hambleton, P. Merenda y S. Spielberger (Eds.), *Adapting educational and psychological test for cross-cultural assessment* (pp. 3-38). Filadelfia, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. y Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, *1*(1), 1-30.

Hernández, R., Fernández, C. y Baptista, P. (2010). *Metodología de la investigación*. Lima: McGraw-Hill.

Herrera Rojas, A. (1993). *La medición en psicología*. Bogotá: Universidad de Bogotá.

Herrero, J. (2010). El análisis factorial confirmatorio en el estudio de la estructura y estabilidad de los instrumentos de evaluación: Un ejemplo con el cuestionario de autoestima. *Intervención Psicosocial*, *19*(3), 289-300. doi: 10.5093/in2010v19n3a9

Hirschfeld, G. y Von Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R - A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assesment, Research & Evaluation*, *19*(7), 1-12.

Hu, L. y Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Model*, *6*, 1–55. doi: 10.1080/10705519909540118

INEGI. (2013). *Estadísticas*. Recuperado de http://www.inegi.org.mx/sistemas/olap/proyectos/bd/consulta.asp?p=17118&c=27769&s=est

Instituto Nacional para la Evaluación de la Educación. (2007). *Disciplina, violencia y consumo de sustancias nocivas a la salud en escuelas primarias y secundarias de México*. Ciudad de México: Autor.

Instituto Nacional para la Evaluación de la Educación. (2015). *Comunicado de prensa nº1*. Ciudad de México: Autor.

Instituto Nacional para la Evaluación de la Educación. (2015). *Plan Nacional para las Evaluaciones de los Aprendizajes*. Ciudad de México: Autor.

International Test Commission. (2010). *International Test Commission guidelines for translating and adapting tests*. Nueva York, NY: International Test Commission.

Jiménez, J. F. (2000). *Derechos de los niños. Colección nuestros derechos*. Ciudad de México: Instituto de Investigaciones Jurídicas.

Jornet, J., González-Such, J. y Perales, M. (2012). Diseño de cuestionarios de contextos para la evaluación de sistemas educativos: Optimización de la medida de constructos complejos. *Bordón, 64*(2), 89-108.

Kim, K. y Bentler, P. (2006). Data modeling: Structural equation modeling. En J. Green, G. Camili y P. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 161-175). Nueva York, NY: American Education Association.

Klerk, G. (2008). *Cross-cultural testing. In online readings in testing and assessment*. Recuperado de http://www.intestcom.org/Publications/ORTA.php

Luo, H. (2011). *Some aspects in confirmatory factor analysis of ordinal variables and generating non-normal data*. Recuperado de http://www.divaportal.org/smash/record.jsf?pid=diva2:405108

López-González, E., Tourón, J. y Tejedor, F. (2012). Diseño de un micro-instrumento para medir el clima de aprendizaje en cuestionarios de contexto. *Bordón, 64*(2), 111-126.

Magnus, L. (2005). Examining the validity of a Swedish version of the self-presentation in exercise questionnaire. *Measurement in Physical Education an Exercise Science, 9*(2), 113-134. doi: 10.1207/s15327841mpee0902_3

Magnusson, D. (1982). *Teoría de los test*. Ciudad de México: Trillas.

Martínez, F. (2013). El futuro de la evaluación educativa. *Sinéctica, 40*. Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-109X2013000100006

Martínez-Arias, R. (1995). *Psicometría: Test de los test psicológicos y educativos*. Madrid: Síntesis.

Messick, S. (1980). Test validity and ethics of assessment. *American psychologist, 35*, 1012-1027.

México. (1982). *Acuerdo Secretarial 96 que establece la organización y funcionamiento de las escuelas primarias*. Diario Oficial de la Federación, de 7 de diciembre de 1982.

México. (1982). *Acuerdo Secretarial 97 que establece la organización y funcionamiento de las Escuelas Secundarias Técnicas*. Diario Oficial de la Federación, de 3 de diciembre de 1982.

México. (1982). *Acuerdo Secretarial 98 por el que se establece la organización y funcionamiento de las Escuelas de Educación Secundaria*. Diario Oficial de la Federación, de 7 de diciembre de 1982.

México. (1993). *Ley General de Educación*. Diario Oficial (Separata), de 13 de julio de 1993.

México. (2002). *Ley para la protección de los derechos de niñas, niños y adolescentes*. Diario Oficial de la Federación, núm. 19, de 29 de mayo de 2000, pp. 2-10.

México. (2013). *Acuerdo 705 por el que se emiten las Reglas de Operación del Programa de Escuela Segura*. Diario Oficial de la Federación, de 28 de diciembre de 2013, pp. 52-71.

Muñiz, J. y Hambleton, R. (1996). Directrices para la adaptación y traducción de instrumentos. *Papeles del Psicólogo, 66*, 63-70.

Muñiz, J., Elosua, P. y Hambleton, R. (2013). Directrices para la traducción y adaptación de los test: Segunda edición. *Psicothema, 25*(2), 151-157. doi: 10.7334/psicothema2013.24

OCDE. (2011). *Establecimiento de un marco para la evaluación de incentivos docenes: Consideraciones para México*. Ciudad de México: Autor.

Ochoa, A. y Salinas, J. (2013). Diagnóstico de la convivencia escolar en escuelas de educación básica de la ciudad de Querétaro. En R. Hevia y J. S. Bravo (Orgs.), *Actas del V Congreso Iberoamericano de Violencia Escolar. Conversar la cultura escolar para construir convivencia (*pp. 56-75). Santiago de Chile: UDP - OEI - OVE.

Ortega-Ruiz, R., Del Rey, R. y Casas, J. A. (2013). La convivencia escolar: Clave en la predicción del bullying. *Revista Iberoamericana de Evaluación Educativa, 6*(2), 91-102.

Pickering, M. y Plotnikoff, R. (2009). Factor structure and measurement invariance of 10-item decisional balance scale: Longitudinal and subgroup examination within an adult diabetic sample. *Measurement in Physical Education and Exercise Science, 13*(4), 206-226. doi: 10.1080/10913670903260086

Ponce, V. (2009). Investigación y políticas educativas. *Revista Electrónica Sinéctica, 33*, 1-33.

Raju, N., Laffitte, L. y Byrne, B. (2002). Measurement equivalence a comparison methods based on confirmatory analysis and item response theory. *Jounal Applied Psychology, 87*(3), 517-531. doi: 10.1037//0021-9010.87.3.517

Ramada, J., Serra, C. y Delclós, G. (2013). Adaptación cultural y validación de cuestionarios de salud: Revisión y recomendaciones metodológicas. *Salud Pública de México, 55*(1), 57-66.

Ramos, R. Y. (2001). *Educacion integral. Una educación holística para el siglo XXI*. Madrid: Desclée de Brouwer.

Reyes, E. P. (2014). *Validez del cuestionario de opinión de alumnos universitarios sobre la competencia docente* (Tesis doctoral). Universidad Autónoma de Baja California, México.

Rodríguez, J. C. (en prensa). *Reporte técnico: Estrategia evaluativa 2015. Factores asociados al aprendizaje*. Ensenada: UEE-UABC.

Rossi, P. y Freeman, H. (1993). *Evaluation. A systemic approach*. Londres: Sage.

Secretaría de Educación Pública. (2007). *Programa Nacional Escuela Segura*. Ciudad de México: Autor.

Secretaría de Educación Pública. (2011). *Programas de Estudio 2011. Guía para el maestro. Educación Básica. Secundaria. Formación Cívica y ética*. Ciudad de México: Dirección General de Desarrollo Curricular-Subsecretaría de Educación Básica.

Secretaría de Educación Pública. (2013). *Programa Sectorial de Educación 2013-2018*. Ciudad de México: Autor.

Silva, A. (1992). *Métodos cuantitativos en psicología*. Ciudad de México: Trillas.

Sireci, S., Patsula, L. y Hambleton, R. (2005). Statistical methods for identifying flaws in the test adaptation process. En R. Hambleton, P. Merenda y S. Spielberger (Eds.), *Adapting educational and psychological test for cross-cultural assessment* (pp. 93-115). Filadelfia, NJ: Lawrence Erlbaum Associates.

Tanzer, N. K. y Sim, C. Q. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC Guidelines for Test Adaptations. *European Journal of Psychological Assessment, 15*(3), 258–269. doi: 10.1027//1015-5759.15.3.258

Valadez, I. (2008). *Violencia escolar: Maltrato entre iguales en escuelas secundarias de la zona metropolitana de Guadalajara.* Ciudad de México: Mar-Eva.

Van de Vijver, F. y Hambleton, R. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2), 89–99. doi: 10.1027/1016-9040.1.2.89

Van de Vijver, F. y Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. En R. Hambleton, P. Merenda y S. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Filadepfia, NJ: Lawrence Erlbaum Associates.

Villalobos, F., Arévalo, C. y Rojas, F. (2012). Adaptación del Inventario de Resilencia ante el Suicidio (SRI-25) en adolescentes y jóvenes de Colombia. *Revista Panamericana de Salud Pública, 31*(3), 233-239. doi: 10.1590/S1020-49892012000300008

Breve CV de los autores

Cristina Vanessa Hernández De la Toba

Doctora en Ciencias Educativas en el Instituto de Investigación y Desarrollo Educativo (IIDE) de la Universidad Autónoma de Baja California (UABC). Directivo-administrativo en el ámbito escolar en áreas de Planeación Institucional, Desarrollo Curricular, Servicios Escolares y de Subdirección. Profesora de educación secundaria, preparatoria y educación superior. Investigadora en Evaluación Educativa. Investigaciones recientes: *Desarrollo y validación de pruebas criteriales.* ORCID ID: 0000-0002-0264-1296. Email: cristina.vanessa.hernandez.delatoba@uabc.edu.mx

Joaquín Caso Niebla

Doctor en Psicología Educativa por la Universidad Nacional Autónoma de México (UNAM). Investigador del Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California, en el área de evaluación educativa. Miembro del Sistema Nacional de Investigadores. Investigaciones recientes: *Desarrollo y validación de instrumentos de medición psicológica y educativa, Evaluación del aprendizaje de estudiantes de educación básica, Variables asociadas al rendimiento académico de adolescentes mexicanos*, y *Evaluación de la convivencia escolar.* Actualmente se desempeña como Director del Instituto de Investigación y Desarrollo Educativo de la UABC. ORCID ID: 0000-0002-3557-1722. Email: jcaso@uabc.edu.mx

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 153-173.
https://doi.org/10.15366/riee2017.10.1.008

Diseño de un Instrumento para Evaluar el Valor Social Subjetivo de la Educación en Estudiantes, Docentes y Familias: Resultados de un Ensayo Piloto

Design of an Instrument to Assess the Subjective Social Value of Education in Students, Teachers and Families: Results of a Pilot Study

Carlos Sancho-Álvarez *
Jesús Miguel Jornet Meliá
José González-Such

Universitat de València

Encontramos multitud de estudios sobre el diseño y validación de instrumentos de evaluación sobre variables socio-afectivas complejas. En estos casos ha sido evidente que un desarrollo sistemático de estandarización ha mejorado las propiedades métricas y su adecuación. Por ello planteamos algunos resultados previos sobre un estudio piloto para la construcción de un instrumento que evalúe el Valor Social Subjetivo de la Educación en tres audiencias (estudiantes, profesorado y familias). En este caso se presentan las propiedades métricas encontradas y las diferencias entre grupos con 131 estudiantes, 28 docentes y 36 familiares. En general los datos son adecuados en las tres escalas implementadas, sin embargo se encuentran nuevas líneas de indagación. Como conclusión se reflexiona sobre la necesidad de asegurar la fiabilidad y la validez de la escala para poder mejorar la investigación.

Palabras clave: Evaluación educativa, Indicadores educativos, Metodología de investigación, Medición, Estudiantes, Docentes, Familias.

We found many studies about the design and validation of evaluation instruments with complex socio-affective variables. In these cases, it has been clear that a systematic scheme of standardization has improved the metric properties and their suitability. Therefore, we propose some preliminary results of a pilot study for the construction of an instrument to assess the Subjective Social Value of Education in three audiences (students, teachers and families). In this case, are presented the metric properties found and the differences between groups with 131 students, 28 teachers and 36 family members. In general, in the in the three scales implemented the dates are suitable, however we found new lines of inquiry. In conclusion, we reflect about the need to ensure the reliability and the validity of test for we can to improve the study.

Keywords: Educational assessment, Educational indicators, Research methodology, Measurement, Students, Teachers, Families.

*Contacto: carlos.sancho@uv.es
issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 2 de septiembre de 2016
1ª Evaluación: 18 de diciembre de 2016
Aceptado: 16 de febrero de 2016

1. Introducción[1]

Resulta habitual ya en investigación la relación que se establece entre el nivel socio-económico y cultural familiar y el rendimiento académico. Son muchos los estudios que han corroborado esta asociación de factores desde hace décadas (Coleman et al., 1966; Casey et al., 2011; Gamoran y Long, 2006; Heyneman y Loxley, 1983; Jeynes, 2002).

Gracias a otras investigaciones se está observando que esta relación va perdiendo consistencia si observamos otras características contextuales sobre los resultados escolares del alumnado. Tal y como señalan Joaristi, Lizasoain y Gamboa (2011):

> *La investigación señala que se da un importante efecto contextual asociado a las características demográficas del aula o del centro escolar; efecto que tiene o puede tener una influencia incluso superior a los efectos individuales a nivel familiar (Brookover et al., 1978; Henderson, Mieszkowski y Sauvageau, 1978; Rumberger y Willms, 1992; Shavit y Williams, 1985; Willms, 1986; Tajalli y Opheim, 2004; Howley y Howley, 2004; Warschauer et al., 2004).* (p. 153)

El contexto (*background*) del alumnado se debería investigar y utilizar los resultados para potenciar o mejorar las diferentes situaciones (Joaristi, Lizasoain y Gamboa, 2012; Jornet, González-Such y Perales, 2012; López-González, González-Such y Lizasoain, 2012; Murillo, 2009). Es decir, sin llegar a anticipar diferencias en los niveles de desempeño en colectivos desfavorecidos simplemente por su nivel de recursos o conocimiento. Algo que también es complejo ya que para poder mejorar una situación primero se debe conocer y comprender. De ahí que el primer objeto de estudio de la sociología de la educación fuera dar soporte a la relación entre origen social y resultados académicos (Carabaña, 2016). Pero volvemos a la misma argumentación anterior; debemos utilizar de manera profesional este tipo de evidencias debido al gran efecto de perversión o incluso de frustradas expectativas que podamos estar creando sin ser conscientes –o tristemente a veces siéndolo–.

En esta línea, como se va observando desde hace varios años –aproximadamente desde 2009– el informe de evaluación internacional de estudiantes PISA ya encuentra serias dificultades para poder explicar un adecuado porcentaje de la varianza entre correlaciones significativas asociadas al estatus socio-económico y cultural de las familias (ECSC, por sus siglas en inglés) y las competencias escolares (OCDE, 2012; 2014). Asimismo, cada vez se estudian más los errores de medida de algunos índices de riqueza familiar (Traynor y Raykov, 2013), o incluso problemas de validez cultural (Solano-Flores, Contreras y Backhoff, 2006).

Por todo ello se ha ido viendo a lo largo de los diferentes informes de evaluación en muchos países que si analizamos el Index of Economic, Social and Cultural Status (ESCS) la fuerte relación que se ha ido corroborando se ve afectada por la superación del logro esperado respecto a algunos colectivos determinados. Por ejemplo, el rendimiento competencial de familias en situaciones favorables es menor al esperado. Así como los

[1] El estudio fue realizado en el marco del proyecto I+D+I "Sistema Educativo y Cohesión Social: diseño de un modelo de evaluación de necesidades (SECS/EVALNEC)", ref. EDU2012-34734 financiado por el Ministerio de Economía y Competitividad (ESPAÑA). Así como con "Ajudes per a la formació de personal investigador de caràcter predoctoral, en el marc del Subprograma Atracció de Talent 2013" del Vicerectorat d'Investigació i Política Científica de la Universitat de València (ESPAÑA).

estudiantes de contextos vulnerables están superando las expectativas en relación a su nivel socio-económico y cultural (Jornet, 2012; Sancho-Álvarez, Jornet y González-Such, 2016), como se puede observar a continuación a través de los siguientes casos.

Para poder ilustrar mejor este tipo de afirmaciones veremos a continuación algunos ejemplos concretos que se vienen dando como apoyo a esta tendencia.

En relación a la variación del rendimiento entre países a nivel internacional en las tres áreas evaluadas por la OCDE (PISA) y el nivel de estudios de la familia podemos observar los gráficos 1 y 2.

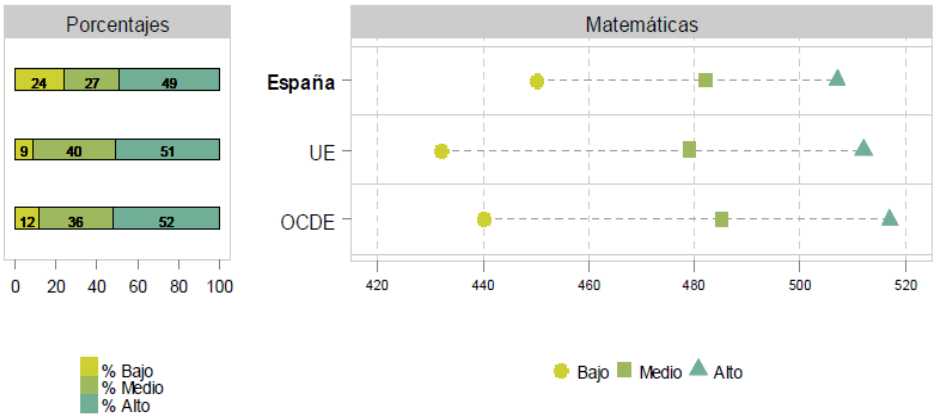


Gráfico 1. Puntuaciones medias y porcentajes en competencia matemática según el nivel de estudios de la familia
Fuente: OECD (2013, p. 93).

Por lo que los estudiantes españoles de familias con estudios bajos y medios obtienen mayores resultados de los esperados, en contraposición al alumnado de familias con estudios superiores que no llega a alcanzar las expectativas esperadas.

Asimismo, si observamos la puntuación media en competencia lectora y en ciencias respecto al nivel de estudios de los padres, la situación aparece tal como se observa en el gráfico 2.

Cabe destacar en relación a la situación del estudiantado en España respecto a la competencia lectora y en ciencias que la situación es similar a la anterior. Ya que el alumnado perteneciente a familias cuyos padres tienen estudios de nivel bajo y medio superan significativamente los resultados esperados de acuerdo al promedio, si la variabilidad fuera la misma asociada a este factor, en relación a la UE y la OECD. Así como los resultados esperados de los estudiantes con padres de niveles de estudios altos no alcanzan la expectativa esperada de logro académico.

En este sentido, podemos observar también la relación en la variación del rendimiento entre países a nivel internacional en las tres áreas evaluadas por la OECD (PISA) y el nivel de ocupación de la familia –ver gráfico 3–.

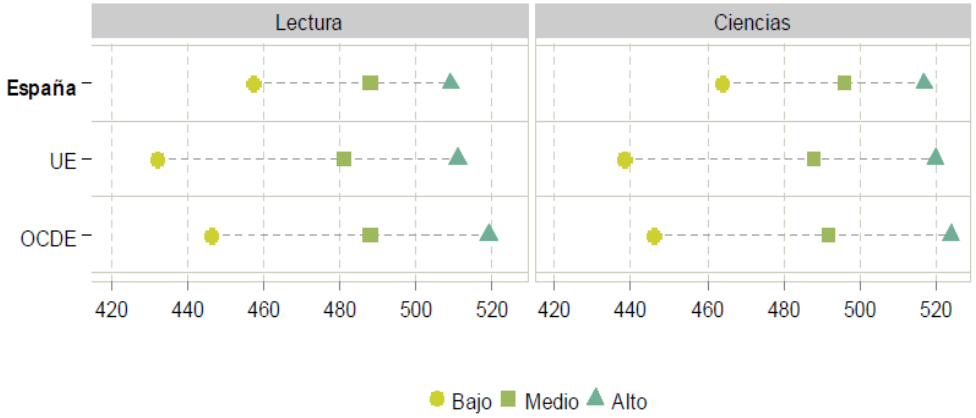


Gráfico 2. Puntuaciones medias y porcentajes en competencia lectora y en ciencias según el nivel de estudios de la familia
Fuente: OECD (2013, p. 94).

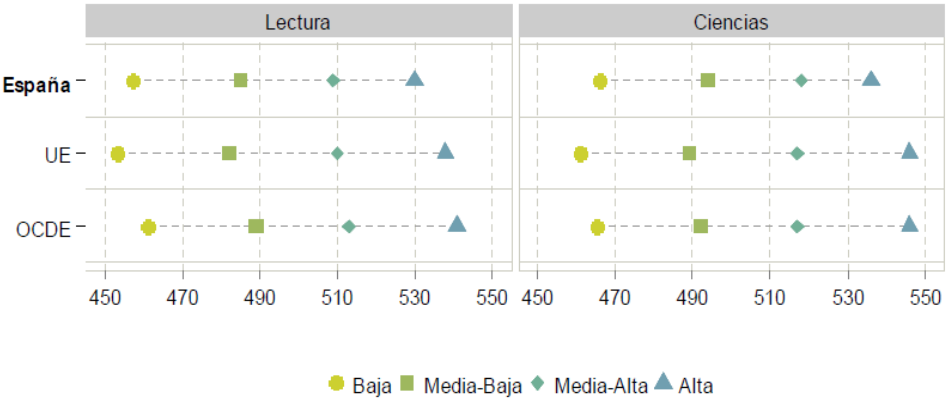


Gráfico 3. Puntuaciones medias y porcentajes en competencia lectora y en ciencias según el nivel de ocupación de la familia
Fuente: OECD (2013, p. 96).

En el caso de la competencia lectora y en ciencias continúa la tendencia en los niveles más bajos de ocupación familiar de superar las expectativas de logro, mientras que con respecto a los niveles de empleo más alto de la familia se mantiene o baja el logro.

Tal es la paradoja y compleja relación que, en esta línea argumentativa, también encontramos otras evidencias en relación a la asociación entre provincias y resultados por ejemplo en el caso del estado español, donde podemos comprobar que existe una clara diferencia entre norte y sur, si observamos las diferencias que hemos ido analizando anteriormente (Gil, 2014) a nivel internacional (OECD, 2013).

Considerando que la relación entre el ESEC y el nivel de logro parece que no está tan clara como ocurría en décadas anteriores, entendemos que pueden existir otras variables o constructos que interaccionen con ello, propiciando diferencias en el desempeño esperado. Entre ellas, pueden posiblemente identificarse muchas, si bien, en el modelo de Evaluación de sistemas basado en el concepto de Cohesión Social, hemos venido

desarrollando el análisis de posibles constructos no habituales[2]. Entre ellos, en este caso, nos referimos al valor que la sociedad en general y las personas, en particular, dan a la educación. El supuesto que sustenta este estudio es que si las personas perciben la educación como un elemento de transformación y desarrollo personal y social es posible que ello facilite el incremento de la motivación y comprometa al alumnado en el esfuerzo por mejorar. Orientadores y psicopedagogos tienen claro, por su experiencia, que las expectativas que se crean en la familia y el profesorado, si son interiorizadas por el alumnado, favorecen mejores rendimientos (Ferrández-Berrueco y Sánchez-Tarazaga, 2014, López y Pantoja, 2016).

De hecho, en los informes de la OCDE *Education at a Glance*, en todas sus ediciones, se incluyen indicadores acerca del beneficio diferencial que tiene la educación. En el gráfico 4 ponemos tan sólo un ejemplo. Probablemente, si este tipo de informaciones llegara, a través de los procesos de orientación al alumnado y sus familias, se podría generalizar la idea de que, en cualquier caso, la educación siempre conlleva beneficios y previene de la exclusión social (Julià, Escapa y Marí-Lose, 2015). Desde esta perspectiva propusimos como constructo la Percepción del Valor Social de la Educación, como un factor a considerar en la evaluación de sistemas e instituciones educativas y, como un elemento importante para la explicación del desempeño. Sobre su definición, nos ocupamos a continuación.

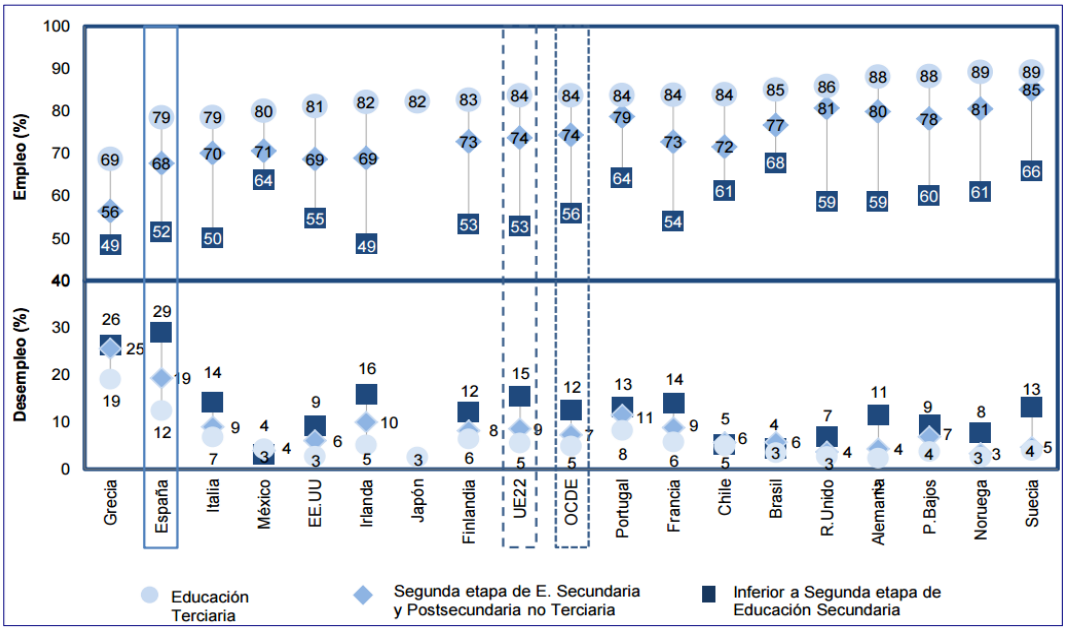


Gráfico 4. Porcentajes de empleo y desempleo según países y etapa educativa estudiada
Fuente: Instituto Nacional de Evaluación Educativa (2016, p. 32).

[2] Por ejemplo el constructo de Justicia Social, Resiliencia o la Competencia Emocional. Variables que se deben seguir estudiando ya que podrían estar influyendo directamente en el rendimiento académico del estudiante (De Pedro y Muñoz, 2005; Hidalgo y Murillo, 2016; Jornet, Sancho-Álvarez y Bakieva, 2015; Marchant, Milicic y Alamos, 2015).

2. El Valor social de la educación

Si bien es cierto que el contexto puede influir en el rendimiento del estudiante, aspectos objetivos como la propia inversión de un gobierno en educación o el sueldo de un maestro, (sin duda la visión subjetiva sobre ese contexto), puede crear una expectativa e importancia dentro del proceso educativo que también influya en cierta manera en la implicación del alumnado en sus estudios. Esta serie de aspectos, sin importar la cultura o el nivel económico de origen, son verdaderamente relevantes para poder avanzar en equidad y mejorar el nivel de justicia social de todas las familias. En este sentido se sitúa el presente estudio donde el principal objetivo es aportar un instrumento de evaluación que pueda discriminar el VSE-Subjetivo para determinar aspectos contextuales que pueden estar influyendo de manera notoria sobre la práctica educativa.

Se considera necesaria la evaluación del proceso escolar desde un modelo que pueda enfocar la enseñanza-aprendizaje hacia la cohesión social, previniendo y mejorando diferentes problemáticas socio-educativas. Así entendemos que la propuesta de Jornet (2012) puede ser adecuada para poder discriminar diferentes factores contextuales sobre la práctica escolar. También de acuerdo con Sancho-Álvarez, Jornet y González-Such (2016) resultan relevantes diferentes trabajos sobre validación del constructo VSE-Subjetivo en esta línea, ya que se han obtenido diferentes evidencias de validación por jueces, su operativización en dimensiones y respecto al diseño y formulación de ítems.

En este contexto nos referimos al concepto de *Valor Social de la Educación* (VSE, en adelante). De acuerdo con Jornet, Perales y Sánchez-Delgado (2011) el VSE se refiere a:

> *La utilidad que tiene la educación dentro de una sociedad para el desarrollo y la promoción de las personas en los ámbitos social y laboral, así como a las ventajas que aporte como elemento de prevención de la exclusión social, y como garantía para el desarrollo y la mejora de su bienestar a lo largo de la vida.* (p. 53)

En el propio concepto de VSE se encuentra inmerso el *Valor Social Subjetivo de la Educación* (VSE-Subjetivo, en adelante) como una dimensión que se construye por "la percepción que los actores principales del proceso de enseñanza-aprendizaje (alumnado, familias y profesorado) tienen acerca de la importancia de la educación para la promoción social, laboral y del bienestar personal y colectivo, a lo largo de la vida" (p. 67). Es decir, en definitiva, se trata de conocer si en los diversos actores del proceso educativo se asume que la educación puede considerarse un "ascensor para la promoción personal y social" o no.

3. Metodología

Se trata de un estudio dirigido al diseño de una escala que permita medir el Valor Social-Subjetivo de la Educación en alumnado, profesorado y familias. Como tal, es pues una investigación de tipo diferencial-correlacional, orientada al diseño de instrumentos de medición. El Objetivo Principal es la depuración de la Escala a partir de un ensayo piloto, de carácter empírico. Como Objetivos Específicos incluimos el análisis de propiedades métricas de la Escala y de los ítems, según el modelo de Consistencia Interna (α de Cronbach) y un análisis diferencial en los subgrupos de participantes en relación a variables demográficas.

Por lo tanto, el foco de atención de este estudio son los resultados previos obtenidos para facilitar una primera revisión de criterios de bondad de las escalas, proceder a su depuración basada en criterios métricos y proponer un modelo de instrumento que pueda posteriormente ser ya tratado en términos de estudios orientados a la generalización de su eficacia, eficiencia y funcionalidad.

A partir de este marco, consideramos pertinente y de utilidad explicativa del desempeño desarrollar un instrumento de evaluación sobre el VSE-Subjetivo. El objetivo general es validar escalas diferenciadas para distintas audiencias, para estudiantes, docentes y familias dentro de un marco de evaluación hacia la Cohesión Social (Jornet, 2012). Asimismo, como objetivos específicos, el presente estudio se centra en: (1) analizar las propiedades métricas del instrumento (de acuerdo a cada escala) mediante la Teoría Clásica del Test (TCT), tomando como referencia el Modelo de Consistencia Interna (α de Cronbach), (2) identificar los ítems defectuosos para mejorar el funcionamiento métrico de las escalas[3], así como (3) analizar por medio de contrastes de hipótesis (en este caso no-paramétricos) si existen diferencias estadísticamente significativas entre algunas variables socio-demográficas en el comportamiento del constructo analizado para cada caso. De manera complementaria se realiza un análisis de componentes principales categóricos –CATPCA– para estudiantes, con el fin de analizar si existe una asociación de carácter multivariado entre los elementos/dimensiones de las escalas y los colectivos a que se dirigen.

4. Instrumentos

Se han administrado tres escalas diferenciadas[4] para cada audiencia compuestas cada una de ellas de 20 ítems distribuidos en cuatro dimensiones teóricas. El constructo teórico de VSE-Subjetivo se compone de la Dimensión 1: Expectativas y metas escolares; Dimensión 2: Valor diferencial de la Educación; Dimensión 3: Justicia Social y Educación y la Dimensión 4: Obstáculos y facilitadores. En las cuales se distribuyen de acuerdo a la dimensión 1 cuatro ítems (1.1-1.4), la dimensión 2 siete ítems (2.1-2-7), la dimensión 3 cinco ítems, (3.1-3.5) y la dimensión 4 cuatro ítems (4.1-4.4) (Sancho-Álvarez, Jornet y González-Such, 2016).

La formulación de los reactivos se presenta a continuación para cada grupo de estudio, de acuerdo a su orden de administración, en el que se ha considerado la formulación de ítems en cuanto a direcciones positivas y negativas, así como la unificación de enunciados.

[3] Para la depuración y filtro de ítems defectuosos entre las escalas se han seguido las recomendaciones de Morales-Vallejo, Urosa-Sanz y Blanco-Blanco (2003).

[4] Como puede observarse, en esencia, se ha intentado mantener el contenido de los ítems en las tres escalas, formulando cada reactivo en función del colectivo al que se dirige, por lo que las diferencias en cuanto al contenido son mínimas, de manera que se facilite la comparabilidad entre las percepciones de cada grupo.

Tabla 1. Escalas del Valor Social Subjetivo de la Educación

SECS/EVALNEC VSE-SUBJETIVO ESTUDIANTES	SECS/EVALNEC VSE-SUBJETIVO DOCENTES	SECS/EVALNEC VSE-SUBJETIVO FAMILIAS
Estudiar me ayuda a aprobar (3.1)	Estudiar ayuda para aprobar (3.1)	Estudiar ayuda a aprobar (3.1)
Mis profesores/as creen que voy a suspender (1.1)	Creo que mis estudiantes van a aprobar (1.1)	Creo que el profesorado piensa que mis hijos/as van a aprobar (1.1)
Es importante para mí ir bien en los estudios (2.1)	Es importante para mí que el alumnado vaya bien en los estudios (2.1)	Es importante para mí que mis hijos/as vayan bien en los estudios (2.1)
Tengo compañeros/as que aprueban sin merecérselo (3.2)	Tengo estudiantes que aprueban sin merecérselo (3.2)	Hay estudiantes que aprueban sin merecérselo (3.2)
Mis padres quieren que trabaje porque estudiar no sirve para tener más dinero (1.2)	Las familias quieren que sus hijo/as trabajen porque estudiar no asegura que puedan tener mejor trabajo o ganar más dinero en el futuro (1.2)	Quiero que mis hijos/as trabajen porque estudiar no asegura que puedan tener mejor trabajo o ganar más dinero (1.2)
Mis padres creen que es más importante estudiar que ganar dinero (2.2)	Las familias creen que es más importante estudiar que ganar dinero (2.2)	Creo que es más importante para mis hijos/as estudiar que ganar dinero (2.2)
Mis padres quieren que trabaje cuanto antes porque hace falta dinero en casa (1.3)	Las familias quieren que sus hijo/as trabajen cuando antes pues hace falta dinero en casa (1.3)	Quiero que mis hijos/as trabajen cuanto antes, pues hace falta dinero en casa (1.3)
Mis padres creen que voy a aprobar (1.4)	Las familias creen que sus hijos/as van a aprobar (1.4)	Creo que mis hijos/as van a aprobar (1.4)
En el colegio me ayudan a tener confianza en mí mismo (2.3)	La escuela y el instituto ayudan a tener confianza en sí mismos/as para tomar decisiones (2.3)	La escuela ayuda a mis hijos/as a tener confianza en sí mismos/as para tomar decisiones (2.3)
En el colegio me enseñan cosas útiles para un futuro trabajo (2.4)	La escuela y el instituto enseñan cosas al alumnado que podrían ser útiles en un trabajo (2.4)	La escuela enseña cosas a mis hijos/as que podrían ser útiles en un trabajo (2.4)
En el colegio me ayudan a encontrar amigos/as (2.5)	La escuela y el instituto ayudan al alumnado a encontrar amigos/as (2.5)	La escuela ayuda a mis hijos/as a encontrar amigos/as (2.5)
En el colegio me ayudan a aprender a vivir con los demás (2.6)	La escuela y el instituto ayudan al alumnado a vivir en sociedad (2.6)	La escuela ayuda a mis hijos/as a vivir en sociedad (2.6)
Lo que estoy aprendiendo en el colegio me servirá cuando decida buscar trabajo (2.7)	Lo que están aprendiendo los estudiante servirá cuando decidan buscar trabajo (2.7)	Lo que están aprendiendo los estudiantes servirá cuando decidan buscar trabajo (2.7)
Pienso que las personas que estudian triunfan más (3.3)	Pienso que las personas que estudian, tienen más éxito en la vida que las que no tienen estudios (3.3)	Pienso que las personas que estudian, tienen más éxito en la vida que las que no tienen estudios (3.3)
Cuanta más gente haya estudiado en mi ciudad, mejor para todos/as (3.4)	Cuanta más gente haya estudiado en la comunidad (sociedad), mejores niveles de vida habrá en ella; es decir, la Educación contribuye a hacer un mundo mejor (3.4)	Cuanta más gente haya estudiado en la comunidad (sociedad), mejores niveles de vida habrá en ella; es decir, la Educación contribuye a hacer un mundo mejor (3.4)

Fuente: Adaptado de Sancho-Álvarez, Jornet y González-Such (2016).

Tabla 1. Escalas del Valor Social Subjetivo de la Educación (continuación)

SECS/EVALNEC VSE-SUBJETIVO ESTUDIANTES	SECS/EVALNEC VSE-SUBJETIVO DOCENTES	SECS/EVALNEC VSE-SUBJETIVO FAMILIAS
Ser profesor/a es una profesión muy importante (4.1)	Ser docente es una profesión muy importante (4.1)	Ser profesor/a es una profesión muy importante (4.1)
Las personas que triunfan en la vida no han estudiado (3.5)	Las personas con éxito en la vida saben ganar dinero, aunque no hayan estudiado (3.5)	Las personas con éxito en la vida saben ganar dinero, aunque no hayan estudiado (3.5)
Mis amigos y amigas no quieren estudiar porque piensan que no sirve para nada (4.2)	Los estudiantes no quieren estudiar, pues piensan que no sirve para nada (4.2)	Los estudiantes no quieren estudiar, pues piensan que no sirve para nada (4.2)
Los políticos de mi país hacen muchas cosas para que nuestra educación sea mejor (4.3)	Los políticos de mi país hacen muchas cosas para que nuestra educación sea mejor (4.3)	Los políticos de mi país hacen muchas cosas para que nuestra educación sea mejor (4.3)
Las personas famosas normalmente no han estudiado nada (4.4)	Las personas famosas normalmente no han estudiado nada (4.4)	Las personas famosas normalmente no han estudiado nada (4.4)

Fuente: Adaptado de Sancho-Álvarez, Jornet y González-Such (2016).

5. Participantes

Dado que se trata de un estudio piloto, se han seleccionado tres grupos de estudio, de características similares a los que se pretende orientar el instrumento final. En la aplicación se ha tenido en cuenta el control del nivel de comprensión de los ítems.

La escala de VSE-Subjetivo para alumnado se ha aplicado con 131 estudiantes; 71 de Primaria (52 % niñas) y 60 de Educación Secundaria (55% chicas).

En cuanto a la escala de VSE-Subjetivo para docentes se ha trabajado con 28 docentes (18% mujeres) de Educación Primaria (61%) y Educación Secundaria (39%).

Asimismo, la escala de VSE-Subjetivo para familias se administrado con 36 familiares (72% madres/tutoras) en relación al alumnado que ha participado anteriormente.

6. Resultados

6.1.Análisis descriptivos

Si observamos los resultados a partir de los análisis estadísticos descriptivos realizados sobre las respuestas a cada escala, podemos observar las medidas de tendencia central y dispersión en relación a cada audiencia implicada -ver tabla 2-.

Tabla 2. Estadísticos descriptivos por escala para cada audiencia

	ESTUDIANTES						PROFESORADO			FAMILIAS		
	Primaria			Secundaria			Primaria/Secundaria			Primaria/Secundaria		
ítem	μ	σ	CV	μ	σ	CV	μ	σ	CV	μ	σ	CV
1.1	3,3	0,74	22,4	3,0	1,0	32,3	3,1	0,57	18,23	3,83	0,38	9,87
1.2	3,4	0,77	22,3	3,6	0,8	22,6	3,2	0,50	15,55	3,50	0,91	26,00
1.3	3,3	0,79	23,8	3,1	0,9	29,1	3,07	0,60	19,67	3,53	0,77	21,93
1.4	3,2	0,80	24,6	3,2	0,8	26,8	3,18	0,61	19,25	3,09	0,89	28,90
2.1	3,5	0,61	17,1	3,5	0,9	24,8	3,39	0,63	18,55	3,58	0,71	19,78
2.2	3,1	0,84	26,6	2,8	1,0	36,9	3,41	0,57	16,77	3,14	0,59	18,89
2.3	3,1	0,84	26,5	2,9	0,8	27,5	2,96	0,74	25,14	3,37	0,81	23,98
2.4	3,4	0,71	20,8	2,4	1,0	43,8	3,32	0,67	20,18	3,54	0,85	24,07
3.1	3,6	0,56	15,2	2,8	0,9	33,4	3,5	0,64	18,23	3,81	0,62	16,38
3.2	3,3	0,72	21,5	3,0	0,8	27,8	3,11	0,69	22,03	3,50	0,74	21,06
3.3	3,3	0,75	22,6	3,6	0,7	18,2	2,93	0,72	24,44	3,56	0,75	20,96
3.4	3,3	0,70	20,8	2,6	0,9	36,6	3,04	0,74	24,47	3,56	0,65	18,31
3.5	3,2	0,85	26,1	3,2	0,8	25,7	3,11	0,69	22,03	3,23	0,84	26,10
3.6	3,2	0,91	27,8	2,8	0,8	29,8	3,21	0,69	21,37	3,72	0,51	13,79
4.1	3,6	0,59	16,0	3,7	0,7	18,3	3,68	0,61	16,63	3,69	0,62	16,91
4.2	2,6	1,0	38,1	2,7	0,9	31,9	2,48	1,01	40,89	1,91	0,85	44,66
4.3	3,1	0,64	20,0	3,1	0,8	25,3	3,11	0,63	20,23	2,83	0,81	28,66
4.4	3,0	0,71	23,5	2,9	1,0	35,7	2,96	0,69	23,41	3,00	0,69	22,87
4.5	2,9	0,97	32,6	2,0	0,9	43,1	2,79	0,74	26,45	3,61	0,80	22,24
4.6	3,5	0,59	16,8	2,5	0,8	31,2	3,18	0,61	19,25	3,06	0,67	22,03
t	3,2	0,42	12,8	3,0	0,4	12,3	3,13	0,47	15,01	3,36	0,27	8,10

Fuente: Elaboración propia.

Asimismo, para observar los resultados de manera global sobre las medias resultantes para cada dimensión teórica, presentamos la información que se detalla en el gráfico 5.

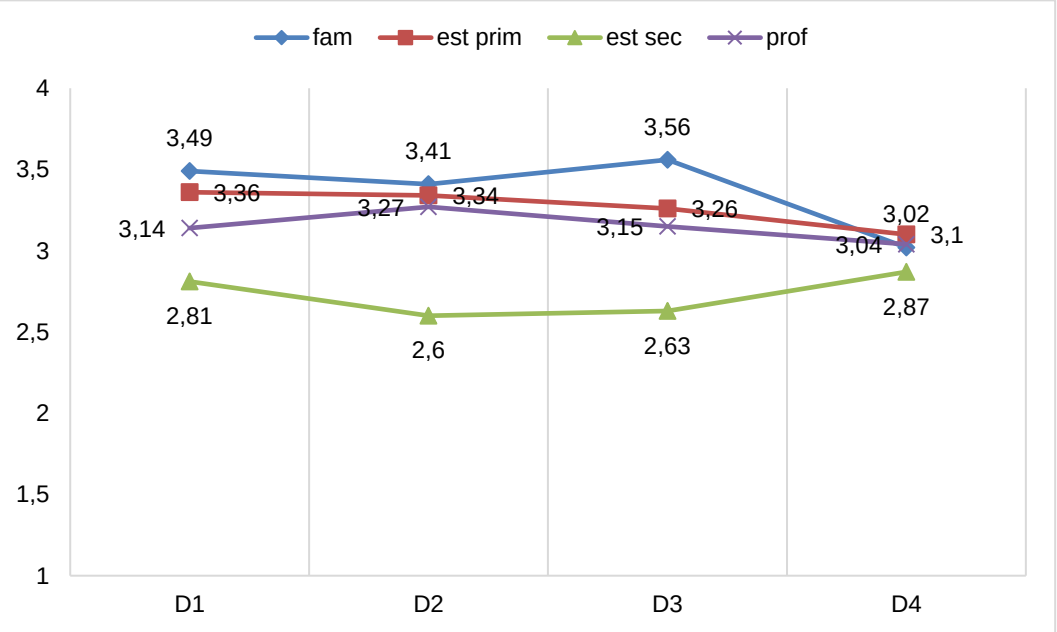


Gráfico 5. Medias por dimensión para escalas estudiantes, profesorado y familias
Fuente: Elaboración propia.

Cabe destacar que el colectivo que presenta las valoraciones más bajas es el alumnado de Educación Secundaria, algo muy diferente que con sus pares de Primaria. Las familias

tienen una percepción alta respecto al constructo, y los docentes unas puntuaciones adecuadas también en todas las dimensiones.

En general, resultan adecuados los datos para cada escala entre dimensiones, ya que la escala es del 1 al 4. Sin duda, la dimensión 4 destaca entre el resto por la cercanía global que se produce entre los colectivos estudiados. Asimismo, señalar que, en conjunto, la mayoría de los reactivos, teniendo en cuenta su cociente de variación, presentan bastante homogeneidad en las respuestas que ofrece cada colectivo (salvo algunos ítems a los que nos referiremos en el apartado de discusión).

6.2. Análisis de fiabilidad

Después de realizar los análisis estadísticos pertinentes en relación a las propiedades métricas de la escala de VSE-Subjetivo para estudiantes, se observa un alfa de Cronbach de 0,750 calculada a partir de 20 reactivos para el total del grupo (Primaria y Secundaria). De manera diferencial para el grupo de Primaria alfa es de 0,570 y de 0,694 para el de Secundaria. En este sentido, presentamos las propiedades métricas de cada ítem –ver tabla 3–.

Tabla 3. Resultados fiabilidad sobre escala estudiantes

ÍTEM	μ			VARIANZA DE ESCALA SI EL ELEMENTO SE HA SUPRIMIDO			ALFA DE CRONBACH SI EL ELEMENTO SE HA SUPRIMIDO		
	t	Prim	Secun	t	Prim	Secun	t	Prim	Secun
1.1	57,56	62	51,88	68,14	38,67	48,76	0,75	0,55	0,68
1.2	57,51	61,67	52,18	66,54	42,16	47,62	0,74	0,59	0,68
1.3	57,44	61,7	51,98	68,23	41,16	50,27	0,74	0,58	0,69
1.4	57,38	61,59	51,98	65,87	40,85	46,43	0,73	0,56	0,66
2.1	57,43	61,58	52,12	63,84	36,85	48,60	0,73	0,52	0,69
2.2	58,03	62,42	52,4	68,70	40,50	48,82	0,75	0,57	0,68
2.3	57,48	61,75	52,02	67,39	38,86	51,20	0,74	0,54	0,69
2.4	57,61	61,61	52,48	65,75	39,96	52,50	0,73	0,55	0,70
2.5	57,41	61,53	52,14	67,11	41,52	50,86	0,74	0,57	0,69
2.6	57,32	61,61	51,82	65,56	37,39	48,23	0,73	0,52	0,67
2.7	57,51	61,69	52,16	63,37	35,74	48,18	0,72	0,50	0,68
3.1	57,77	62,05	52,3	65,49	37,25	48,70	0,73	0,53	0,67
3.2	57,46	61,67	52,06	65,99	40,57	47,08	0,73	0,56	0,66
3.3	57,71	62,05	52,16	65,50	36,62	47,97	0,73	0,53	0,67
3.4	57,27	62,06	51,14	73,86	39,62	51,06	0,77	0,57	0,68
3.5	57,77	62,25	52,04	70,62	41,52	49,75	0,75	0,58	0,68
4.1	57,62	62,23	51,72	71,28	39,67	50,04	0,76	0,56	0,68
4.2	57,57	62,05	51,84	71,31	40,71	52,42	0,76	0,57	0,70
4.3	57,77	61,83	52,58	67,43	42,40	52,00	0,74	0,58	0,70
4.4	57,55	61,73	52,2	69,61	42,64	53,63	0,75	0,59	0,70

Fuente: Elaboración propia.

Como se puede observar en los resultados de fiabilidad para la escala de estudiantes total, se podría incrementar α con la eliminación del ítem 3.4 (Cuanta más gente haya estudiado en mi ciudad, mejor para todos/as), 4.1 (Ser profesor/a es una profesión muy importante) y 4.2 (Mis amigos y amigas no quieren estudiar porque piensan que no sirve para nada). Sin embargo, se decide no eliminarlos porque afectaría a la validez de contenido y además el aumento no es sustancial.

En cuanto a la fiabilidad para el grupo de estudiantes de Primaria se podría aumentar si eliminamos el 1.2 (Mis padres quieren que trabaje porque estudiar no sirve para tener

más dinero), 1.3 (Mis padres quieren que trabaje cuanto antes porque hace falta dinero en casa), 3.5 (Las personas que triunfan en la vida no han estudiado), 4.3 (Los políticos de mi país hacen muchas cosas para que nuestra educación sea mejor) y 4.4 (Las personas famosas normalmente no han estudiado nada). El ítem 1.2 y el 4.4 se podrían eliminar sin que afectara a la validez de contenido, pero se decide mantenerlos porque el aumento no es sustancial. Sin embargo, la eliminación de los otros reactivos sí que afectaría a la validez de contenido por lo que se mantienen.

Al respecto de la escala para el grupo de Secundaria, se podría aumentar la fiabilidad eliminando el 2.4 (En el colegio me enseñan cosas útiles para un futuro trabajo), 4.2, 4.3 y 4.4. Al ser un cambio muy débil en el aumento de la fiabilidad (en todos los casos, el incremento sería de 0,006) y para asegurar la validez de contenido en cada dimensión y por total, se decide mantener todos los ítems.

Por otro lado, a partir de realizar los análisis estadísticos pertinentes en relación a las propiedades métricas de la escala de VSE-Subjetivo para docentes, resulta un alfa de Cronbach de 0,951 a partir de 20 reactivos –ver tabla 4-.

Tabla 4. Resultados fiabilidad sobre escala profesorado

ÍTEM	μ	VARIANZA DE ESCALA SI EL ELEMENTO SE HA SUPRIMIDO	CORRELACIÓN TOTAL DE ELEMENTOS CORREGIDA	ALFA DE CRONBACH SI EL ELEMENTO SE HA SUPRIMIDO
1.1	60,15	85,90	0,75	0,948
1.2	60,04	86,68	0,78	0,948
1.3	60,15	85,34	0,81	0,947
1.4	60,08	85,59	0,72	0,948
2.1	59,88	86,03	0,68	0,949
2.2	59,88	87,87	0,58	0,950
2.3	60,27	84,05	0,72	0,948
2.4	59,92	84,87	0,72	0,948
3.1	59,73	87,57	0,60	0,950
3.2	60,12	84,99	0,72	0,948
3.3	60,27	85,01	0,70	0,948
3.4	60,19	82,72	0,83	0,946
3.5	60,12	84,43	0,77	0,947
3.6	60,00	86,80	0,58	0,950
4.1	59,62	87,69	0,54	0,951
4.2	60,73	81,49	0,67	0,950
4.3	60,15	84,22	0,82	0,947
4.4	60,31	87,10	0,51	0,951
4.5	60,42	84,65	0,69	0,949
4.6	60,08	86,15	0,67	0,949

Fuente: Elaboración propia.

Como se puede observar en los resultados obtenemos una elevada fiabilidad, por lo que finalmente se obtenemos una fiabilidad total de 0,951 sobre la escala compuesta por 20 ítems.

Después de realizar los análisis estadísticos pertinentes en relación a las propiedades métricas de la escala de VSE-Subjetivo para familias, resulta un alfa de Cronbach de 0,676 calculada a partir de 20 reactivos. Para ello, presentamos sus propiedades de manera individual para cada ítem –ver tabla 5-.

En este caso, presentamos los resultados para dos aplicaciones del procedimiento, ya que se decide la eliminación después de la primera aplicación el reactivo 1.2 para mejorar la fiabilidad (1.2: Quiero que mis hijos/as trabajen porque estudiar no asegura que puedan tener mejor trabajo o ganar más dinero). Finalmente, después del segundo análisis (habiendo eliminado el elemento mencionado) se decide eliminar también el ítem 4.6 con lo que se incrementa fiabilidad total a 0,73 con 18 reactivos (4.6: Los estudiantes no quieren estudiar, pues piensan que no sirve para nada (4.6); ya que para ambos casos la validez de contenido no se ve afectada y garantiza una mayor fiabilidad directa con la supresión establecida.

Tabla 5. Resultados fiabilidad sobre escala familias

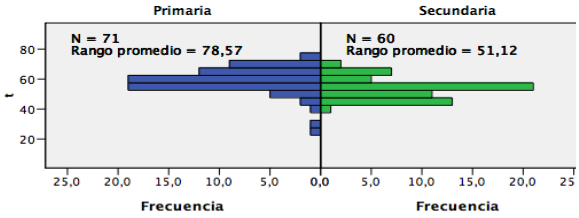
ANÁLISIS 1 ÍTEMS	μ	VAR. SUP	COR REL. TOT	A SUP	ANÁLISIS 2 ÍTEMS	μ	VAR. SUP	COR REL. TOT	A SUP
1.1	63,65	27,36	0,02	0,68	**1.1**	60,12	28,26	-0,01	0,72
1.2	64,00	28,32	-0,17	0,72	**Sup.**	-	-	-	-
1.3	63,88	25,55	0,25	0,67	**1.3**	60,35	26,31	0,24	0,71
1.4	64,54	23,46	0,35	0,65	**1.4**	61,00	24,24	0,34	0,70
2.1	63,96	24,92	0,30	0,66	**2.1**	60,42	25,61	0,31	0,70
2.2	64,27	27,25	0,00	0,69	**2.2**	60,73	27,72	0,05	0,72
2.3	64,19	22,32	0,56	0,62	**2.3**	60,65	22,87	0,58	0,67
2.4	64,04	23,16	0,41	0,64	**2.4**	60,50	23,70	0,43	0,69
3.1	63,62	26,41	0,38	0,66	**3.1**	60,08	27,27	0,34	0,71
3.2	64,00	24,48	0,41	0,65	**3.2**	60,46	24,97	0,45	0,69
3.3	63,88	23,15	0,59	0,63	**3.3**	60,35	23,83	0,59	0,67
3.4	63,96	23,48	0,67	0,63	**3.4**	60,42	24,01	0,70	0,67
3.5	64,42	23,69	0,34	0,65	**3.5**	60,88	24,90	0,28	0,71
3.6	63,77	25,95	0,25	0,67	**3.6**	60,23	26,50	0,29	0,70
4.1	63,85	26,94	0,04	0,68	**4.1**	60,31	27,66	0,04	0,72
4.2	65,69	24,06	0,34	0,65	**4.2**	62,15	24,77	0,34	0,70
4.3	64,62	24,81	0,33	0,66	**4.3**	61,08	25,27	0,37	0,70
4.4	64,58	27,85	-0,11	0,70	**4.4**	61,04	28,19	-0,05	0,73
4.5	63,81	26,16	0,16	0,67	**4.5**	60,27	26,92	0,16	0,71
4.6	64,50	26,26	0,10	0,68	**4.6**	60,96	27,39	0,05	0,73

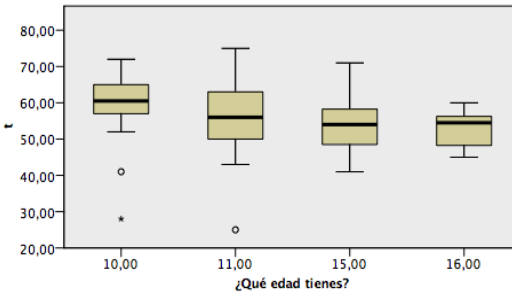
Fuente: Elaboración propia.

6.3. Análisis diferencial entre grupos

6.3.1. Alumnado

A partir de los análisis realizados entre grupos, se obtienen diferencias estadísticamente significativas en la variable etapa (Primaria y Secundaria) y edad (10, 11, 15, 16 años). Tal y como se puede comprobar en los siguientes gráficos.





N total	131
U de Mann-Whitney	1.237,500
W de Wilcoxon	3.067,500
Probar estadística	1.237,500
Error típico	216,249
Estadística de prueba estandarizada	-4,127
Sig. asintótica (prueba de dos caras)	,000

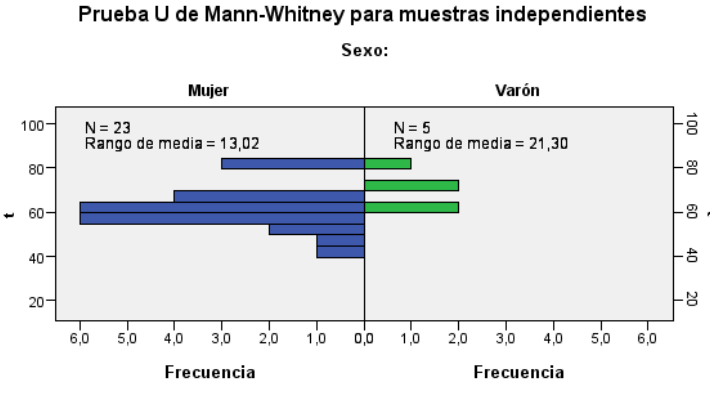
N total	131
Probar estadística	21,575
Grados de libertad	3
Sig. asintótica (prueba de dos caras)	,000

Gráfico 6. Resultados diferencias entre grupos para alumnado
Fuente: Elaboración propia.

Sin embargo, no existen diferencias significativas según otras variables contextuales. Tales como estudios finalizados de mayor nivel de la madre/tutora (p≥ 0,121) o del padre/tutor (p≥ 0,317). Así como tampoco aparecen diferencias estadísticamente significativas entre las categorías de la actividad laboral para la madre/tutora (p≥ 0,760) o sobre el padre/tutor (p≥ 0,340).

6.3.2. Profesorado

Entre los docentes aparecen diferencias estadísticamente significativas en las variables sexo y el nivel de estudios a que aspiran que lleguen sus alumnos.



N total	28
U de Mann-Whitney	91,500
W de Wilcoxon	106,500
Estadístico de contraste	91,500
Error estándar	16,607
Estadístico de contraste estandarizado	2,047
Significación asintótica (prueba bilateral)	,041
Significación exacta (prueba bilateral)	,039

Gráfico 7. Resultados diferencias entre grupo para profesorado
Fuente: Elaboración propia.

No hay diferencias significativas en cuanto a los años de experiencia en la profesión docente (p≥ 0,401) ni tampoco aparecen entre la etapa escolar en la que imparten

actualmente docencia (p≥ 0,578), así como tampoco según hasta qué nivel quieren los docentes que lleguen sus estudiantes (p≥ 0,583).

6.3.3. Familias

No aparecen diferencias estadísticamente significativas en relación al sexo (p≥ 0,689), ni entre edades (p≥ 0,619). Tampoco hay diferencia significativa sobre la variable *¿Hasta qué nivel quieres que tu hijo/a estudie?* (p≥ 0,158). Así como tampoco aparecen diferencias entre el tipo de trabajo que tiene el padre o la madre (p≥ 0,468), ni cuando se pregunta el nivel de estudios finalizados (p≥ 0,767).

6.4. Análisis de componentes principales categóricos CATPCA

A partir de los resultados obtenidos sobre los diferentes análisis creemos adecuado realizar un análisis complementario para la audiencia estudiantes. Ya que, al aparecer diferencias entre edades y etapa puede ser interesante indagar entre estas variables para ver cómo se comportan los datos entre ítems desde una perspectiva multivariada.

Para ello, después de realizar un análisis de componentes principales categóricos CATPCA con las puntuaciones de estudiantes y la variable etapa, se han obtenido dos dimensiones para cada caso –etapa y edad-, siendo el valor propi de la primera 5,48 y 2,07 para la segunda, explicando un 66,3 % de la varianza total –ver su distribución en gráfico 8-.

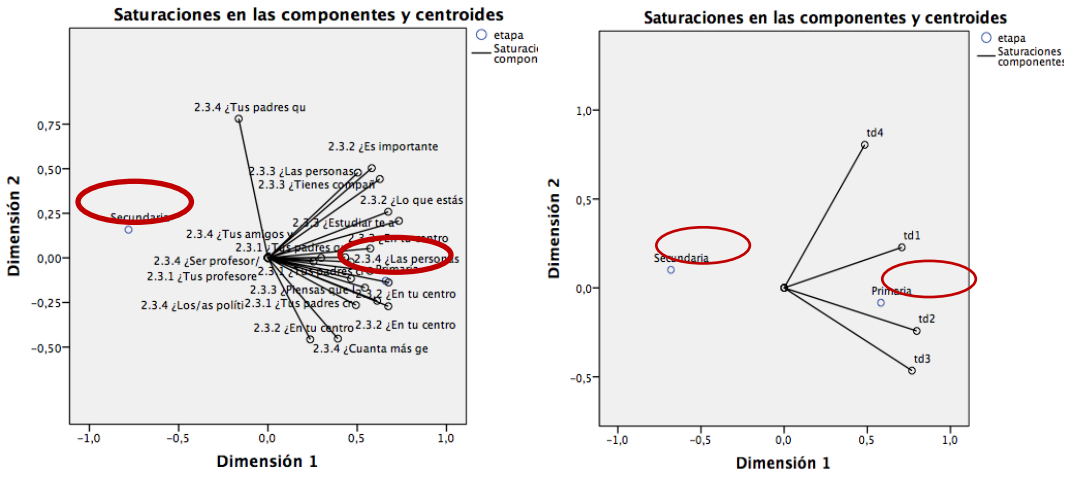


Gráfico 8. Análisis de componentes principales categóricos CATPCA estudiantes etapa
Fuente: Elaboración propia.

Según el gráfico 8, podemos ver el análisis con las valoraciones de estudiantes y la variable etapa educativa, Primaria y Secundaria -a su izquierda para ítems y a su derecha para total por dimensión-. En ambos casos, se puede observar claramente cómo la etapa de educación secundaria se aleja de la de primaria y establece dos componentes claros de saturación. En el caso del grupo de Primaria se relaciona con puntuaciones más altas mientras que el grupo de Secundaria está más cercano a valores cercanos en todos los casos.

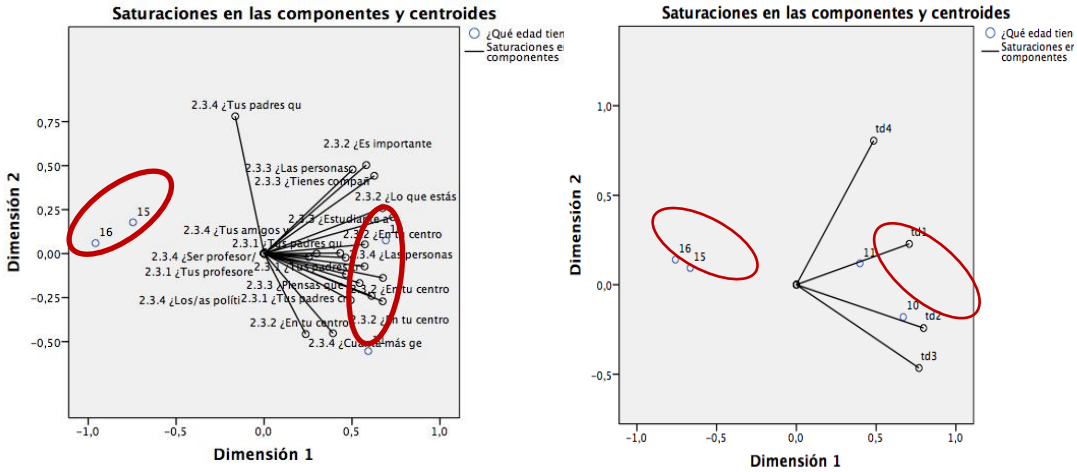


Gráfico 9. Análisis de componentes principales categóricos CATPCA estudiantes edad
Fuente: Elaboración propia.

Según el gráfico 9, así como en el gráfico anterior pero con la variable edad (10 y 11 años de edad y 15 y 16 años de edad). En ambos casos, se puede observar claramente cómo se establecen dos categorías entre las edades de los estudiantes. En el caso del grupo de 10 y 11 años se relaciona con puntuaciones más altas mientras que el grupo de 15 y 16 años está más cercano a valores cercanos en todos los casos.

En general, los análisis vienen a corroborar que el constructo analizado presenta diferencias en la percepción de los estudiantes sobre el VSE-Subjetivo según la etapa y edad en la que se encuentren. Así como, también es coincidente esta diferencia para ambos casos conformando dos categorías diferenciadas.

Todo ello, se tendrá que tener en cuenta para realizar escalas diferenciadas por etapa para investigaciones futuras, para poder dar respuesta adecuadamente a la edad cognitiva.

7. Conclusiones

En relación a los resultados obtenidos sobre criterios de bondad para las tres escalas podemos observar que en general han sido adecuados. Los datos de fiabilidad han sido altos sobre la escala de profesorado, medios en la de estudiantes y moderados-bajos en la de familias. Por esto se justifica la continua revisión de propiedades métricas en futuras investigaciones atendiendo a esas diferencias.

En cuanto a las diferencias entre grupos, vemos que no han aparecido diferencias estadísticamente significativas en las variables estudiadas sobre nivel socio-económico y cultural entre las tres audiencias. Sin embargo, sí que aparecen diferencias entre los estudiantes por etapa y edad. Así como entre los docentes sobre la variable sexo. Lo cual indica una línea de indagación bastante amplia.

En conjunto, analizando el comportamiento de la escala en los diversos colectivos analizados, podemos identificar las siguientes conclusiones:

- Las familias son las que mejor valoración dan a la importancia de la Educación, si bien, la fiabilidad de la escala en este caso es menor que en los demás grupos. Como hipótesis interpretativa, podríamos adelantar que la variabilidad entre familias, sus intereses y expectativas, basadas en sus situaciones socioeconómicas y culturales muy posiblemente sean la base de explicación de las diferencias de opinión.
- Resulta curioso que los estudiantes de Primaria estén más cercanos a las opiniones familiares, otorgando mayor valor a la educación incluso que el profesorado y, que sus compañeros de educación secundaria. En esta etapa evolutiva, el rol de familias y profesorado como modelos de referencia tiene mayor impacto en la percepción y formación de actitudes en los niños que en secundaria.
- El hecho de que los estudiantes de secundaria obtengan los niveles más bajos estimamos que es una evidencia del correcto funcionamiento de la escala, pues en la adolescencia los modelos de referencia pasan a ser preferentemente los del grupo de iguales y es evidente que se trata de una etapa difícil, de construcción de la personalidad, y de establecimiento de metas y objetivos, en la que el alumnado suele tener menor interés por el estudio o, cuanto menos, le reconoce menor valor (Gutiérrez y Expósito, 2015; Julià, Escapa y Marí-Lose, 2015).
- Por otra parte, no encontramos una posible interpretación que nos pueda orientar acerca del motivo de que el profesorado manifieste menor valor de la educación que familias y alumnado de educación primaria. Por ello, estimamos que, en es sin duda una línea prioritaria de investigación posterior, a llevar acabo con el conjunto de la muestra definitiva e indagando sobre posibles evidencias de validación.
- El comportamiento global de la escala en el total y en los subgrupos de Primaria y Secundaria, estimamos que pone de manifiesto diversos aspectos: a) un buen funcionamiento general, siendo escasos los ítems en cada grupo en que su eliminación produciría mejoras en la fiabilidad, si se dan, el incremento es mínimo, b) el hecho de que no sean los mismos ítems los que presentan un funcionamiento deficiente estimamos que refleja un funcionamiento diferencial, propio del carácter evolutivo del constructo, c) en ningún caso, se identifican los mismos elementos con funcionamiento deficiente en los tres grupos: los elementos 4.3. y 4.4. presentan cierto desajuste en Primaria y Secundaria, pero no en el total, y d) la eliminación de alguno de los ítems que, de manera particular en uno u otro grupo presentan pequeñas deficiencias funcionales, puede afectar de manera desigual al mantenimiento de la validez de contenido. Asimismo, los niveles de consistencia interna son muy elevados para el tipo de constructo evaluado. Por ello, se decide mantener la escala tal cual está definida y apuntar como línea de investigación futura la revisión del conjunto de los ítems a nivel evolutivo, a través de cursos/edades en Primaria y Secundaria, con una muestra mayor y analizando su comportamiento tanto por TCT, como por Teoría de Respuesta al Ítem (TRI); y, en definitiva, mantener la escala, con sistemas de baremación diferenciales para curso/edad.

Como líneas posteriores de investigación se plantearía la necesidad de aplicar la Escala en una muestra mayor, para comprobar sus propiedades y observar si las conclusiones planteadas anteriormente se siguen manifestando.

Asimismo, es prioritario realizar estudios sobre evidencias de validación, motivos que nos puedan ir ayudando para avanzar en la investigación e interpretación de hipótesis interpretativas, así como a mejorar definitivamente los procesos de adecuación de la Escala (Ruiz-Primo y Li, 2015; Leyva, 2011).

Referencias

Carabaña, J. (2016). El Informe Coleman, 50 años después. *Revista de la Asociación de Sociología de la Educación*, *9*(1), 9-21.

Casey, B. M., Dearing, E., Vasilyeva, M., Ganley, C. M. y Tine, M. (2011). Spatial and numerical predictors of measurement performance: The moderating effects of community income and gender. *Journal of Educational Psychology*, *103*(2), 296-311. doi:10.1037/a0022516

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M., Weinfeld, F. D. y York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: Government Printing Office.

De Pedro, F. y Muñoz, V. (2005). Educar para la resiliencia: Un cambio de mirada en la prevención de situaciones de riesgo social. *Revista Complutense de Educación*, *16*(1), 107-124.

Ferrández-Berrueco, R. y Sánchez-Tarazaga, L. (2014). Teaching competences in Secondary Education. Analysis of teachers' profiles. *Revista Electrónica de Investigación y Evaluación Educativa (RELIEVE)*, *20*(1), art. 1. doi:10.7203/relieve.20.1.3786

Gamoran, A. y Long, D. A. (2006). *Equality of educational opportunity: A 40 year retrospective.* Madison, WI: WCER.

Gil, J. (2014). Factores asociados a la brecha regional del rendimiento español en la evaluación pisa. *Revista de Investigación Educativa*, *32*(2), 393-410. doi:10.6018/rie.32.2.192441

Gutiérrez, M. y Expósito, J. (2015). Autoconcepto, dificultades interpersonales, habilidades sociales y conductas asertivas en adolescentes. *Revista Española de Orientación y Psicopedagogía*, *26*(2), 42-58. doi:10.5944/reop.vol.26.num.2.2015.15215

Heyneman, S. P. y Loxley, W. A. (1983). The effect of primary school quality on academic achievement across twenty-nine high and low income countries. *American Journal of Sociology*, *88*, 1162-1194.

Hidalgo, N. y Murillo, F. J. (2016). Evaluación de estudiantes para la justicia social. Propuesta de un modelo. *Revista Internacional de Educación para la Justicia Social (RIEJS)*, *5*(2), 159-179. doi:10.15366/riejs2016.5.2.008

Instituto Nacional de Evaluación Educativa. (2016). *Panorama de la educación. Indicadores de la OCDE 2016.* Madrid: Ministerio de Educación, Cultura y Deporte.

Jeynes, W. H. (2002). The challenge of controlling for SES in social science and education research. *Educational Psychology Review*, *14*(2), 205-221.

Joaristi, L., Lizasoain, L. y Gamboa, E. (2011). Construcción y validación de un instrumento de medida del nivel socioeconómico y cultural (NSE) de estudiantes de educación primaria y secundaria. *Bordón*, *64*(2), 151-171.

Joaristi, L., Lizasoain, L. y Gamboa, E. (2012). Construcción y validación de un instrumento de medida del nivel socioeconómico y cultural (NSE) de estudiantes de educación primaria y secundaria. *Bordón*, *64*(2), 151-172.

Jornet, J. M. (2012). Dimensiones docentes y cohesión social: Reflexiones desde la evaluación. *Revista Iberoamericana de Evaluación Educativa*, *5*(1e), 349-362.

Jornet, J. M., González-Such, J. y Perales, M. J. (2012). Diseño de cuestionarios de contexto para la evaluación de sistemas educativos: Optimización de la medida de constructos complejos. *Bordón, 64*(2), 89-110.

Jornet, J. M., Perales, M. J. y Sánchez-Delgado, P. (2011). El valor social de la educación: Entre la subjetividad y la objetividad. Consideraciones teórico-metodológicas para su evaluación. *Revista Iberoamericana de Evaluación Educativa, 4*(1), 51-77.

Jornet, J. M., Sancho-Álvarez, C. y Bakieva, M. (2015). Diseño y validación del constructo teórico de justicia social percibida por el alumnado en los centros escolares. *Revista Internacional de Educación para la Justicia Social (RIEJS), 4*(2), 111-126. doi:10.15366/riejs2015.4.2.

Julià, A., Escapa, S. y Marí-Klose, M. (2015). New social risks and educational vulnerability of boys and girls in Spain. *Revista de Educación, 369*, 9-30. doi:10.4438/1988-592X-RE-2015-369-288

Leyva, Y. E. (2011). Una reseña sobre la validez de constructo de pruebas referidas a criterio. *Perfiles Educativos, 33*(131), 131-154.

López, M. y Pantoja, A. (2016). Diseño y validación de una escala para comprobar la percepción y satisfacción de las familias andaluzas en relación con los procesos tutoriales en centros de educación primaria. *Revista Española de Orientación y Psicopedagogía, 27*(1), 47-66. doi:10.5944/reop.vol.27.num.1.2016.17007.

López-González, E., González-Such, J. y Lizasoain, L. (2012). Explicación del rendimiento a partir del contexto. Algunas propuestas de análisis gráfico y estadístico. *Bordón, 64*(2), 127-149.

Marchant, T., Milicic, N. y Alamos, P. (2015). Competencias socioemocionales: Capacitación de directivos y docentes y su impacto en la autoestima de alumnos de 3º a 7º básico. *Revista Iberoamericana de Evaluación Educativa, 8*(2), 203-218.

Morales-Vallejo, P., Urosa-Sanz, B. y Blanco-Blanco, A. (2003). *Construcción de escalas de actitudes tipo Likert.* Madrid: La Muralla.

Murillo, F. J. (2009). Hacia un modelo de eficacia escolar. Estudio multinivel sobre los factores de eficacia en las escuelas españolas. *Revista Electrónica Latinoamericana sobre Calidad, Eficacia y Cambio en Educación, 6*(1), 4-28.

Organización para la Cooperación y el Desarrollo Económicos. (2012). *PISA 2009. Technical Report.* París: OECD Publishing.

Organización para la Cooperación y el Desarrollo Económicos. (2013). *PISA 2012. Programa para la Evaluación Internacional de los Alumnos: Informe español. Volumen I: Resultados y contexto.* Madrid: MECD.

Organización para la Cooperación y el Desarrollo Económicos. (2014). *PISA 2012. Technical Report.* París: OECD Publishing.

Ruiz-Primo, M. A. y Li, M. (2016). The relationship between item context characteristics and student performance. *Revista Electrónica de Investigación y Evaluación Educativa (RELIEVE), 22*(1). doi:10.7203/relieve.22.1.8280

Sancho-Álvarez, C., Jornet, J. y González-Such, J. (2016). El constructo Valor Social Subjetivo de la Educación: validación cruzada entre profesorado de escuela y universidad. *Revista de Investigación Educativa, 34*(2), 329-350. doi:10.6018/rie.34.2.226131

Solano-Flores, G., Contreras, L. A. y Backhoff, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y

otras comparaciones internacionales. *Revista Electrónica de Investigación Educativa*, *8*(2). Recuperado de http://redie.uabc.mx/vol8no2/contenido-solano2.html

Traynor, A. y Raykov, T. (2013). Household possessions indices as wealth measures: A validity evaluation. *Comparative Education Review*, *57*(4), 662-688. doi:10.1086/671423.

Breve CV de los autores

Carlos Sancho-Álvarez

Máster en Psicopedagogía Social y Comunitaria, y Licenciado en Pedagogía por la Universitat de València, diplomado en Magisterio por la Universidad de Alcalá. Investigador del Departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE) en la Facultad de Filosofía y Ciencias de la Educación, Universitat de València. Miembro del Grupo de Evaluación y Medición GEM-Educo (www.uv.es/gem/gemeduco). Miembro del Grupo de Innovación Docente InnovaMIDE (http://www.uv.es/innovamide). Ha impartido docencia universitaria en Medición educativa y Evaluación de programas en titulaciones de Pedagogía y Educación Social. Actualmente desarrolla su investigación en torno al tema de evaluación de la dimensión del Valor Social de la Educación (y Valor Social Subjetivo de la Educación), en el marco de la evaluación de Sistemas Educativos para la Cohesión Social. ORCID ID: 0000-0001-9489-2502. Email: carlos.sancho@uv.es

Jesús Miguel Jornet Meliá

Doctor en Ciencias de la Educación (1987), Licenciado en Psicología (1979). Profesor de la Universitat de València desde 1984; Profesor Titular de Universidad (1989); Catedrático de Medición y Evaluación Educativas (2006) del área de Métodos de Investigación y Diagnóstico en Educación. Coordinador del GEM (www.uv.es/gem). Actualmente imparte docencia sobre Medición y Evaluación Educativas en diversos Másteres y Doctorados, tanto en la Universitat de València (España), como en otros que se desarrollan en Uruguay, Perú y República Dominicana. Sus líneas de trabajo actuales se centran prioritariamente en: diseño de instrumentos de medición y evaluación educativas (competencias y constructos socio-afectivos, alumnado y profesorado), evaluación de organizaciones y sistemas educativos (dimensión educativa de la cohesión social y equidad como consecuencia de la educación). ORCID ID: 0000-0001-6905-497X. Email: jornet@uv.es

José González-Such

Doctor en Ciencias de la Educación. Profesor Titular de Universidad del área de Métodos de Investigación y Diagnóstico en Educación. Ha impartido docencia sobre Estadística aplicada a las Ciencias Sociales, Medición Educativa, Bases Metodológicas de Investigación Educativa, Técnicas no estandarizadas de recogida de información, Técnicas de medida desde 1987. Profesor de los Másters en Política, Gestión y Dirección de Centros Educativos, Educación Especial y Psicopedagogía de la Universitat de València. Director del máster de Educación Especial de la Universitat de

València. Ha coordinado distintos proyectos de innovación educativa del grupo Innovamide (http://www.uv.es/innovamide). Miembro del Grupo de Evaluación y Medición (http://www.uv.es/gem). Sus principales líneas de investigación son: Medición y Evaluación Educativa, Evaluación del profesorado, Innovación Educativa. ORCID ID: 0000-0001-9086-6446. Email: jose.gonzalez@uv.es

Revista Iberoamericana de Evaluación Educativa

Temática Libre

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 177-193.
https://doi.org/10.15366/riee2017.10.1.009

Proceso General para la Evaluación Formativa del Aprendizaje

General Process for the Formative Assessment of Learning

Eva Pasek de Pinto *[1]
María Teresa Mejía [2]

[1]Universidad Simón Rodríguez
[2]Escuela Bolivariana "Giraluna"

La evaluación formativa se acepta como idónea para mejorar los procesos de enseñanza y aprendizaje. Sin embargo, poco se practica de forma sistemática y aún persiste un enfoque tradicional en muchas escuelas venezolanas al reducirla a una etapa del proceso de aprendizaje necesariamente terminal, enfatizando el producto. En consecuencia, el objetivo de este estudio consistió en configurar un proceso general para la evaluación formativa del aprendizaje a partir de las actividades evaluativas que realizan los docentes en el aula. Metodológicamente se inscribe en el paradigma cualitativo mediante el método etnográfico, el cual se aplicó siguiendo las fases sugeridas por Rodríguez, Gil y García (1999): preparatoria, de campo, analítica e informativa. Para recabar la información se utilizó la observación participante y el diario de campo como instrumento para registrar la información. La información recabada se transcribió, organizó, categorizó y se validó por medio de la triangulación de fuentes. Como resultado se obtuvo un conjunto de actividades de evaluación formativa subyacentes en la práctica pedagógica del docente, las cuales se integraron en una secuencia para conformar un proceso general de evaluación formativa del aprendizaje incluido en los procesos de enseñanza aprendizaje. El proceso es útil para realizar una evaluación formativa consciente y sistemática.

Palabras clave: Evaluación formativa, Investigación, Actividades de evaluación formativa, Proceso general, Educación primaria.

The formative assessment accepted as suitable for improving teaching and learning processes. However, little practiced in a systematic way and still remains a traditional approach in many Venezuelan schools by reducing it to a stage of the learning process necessarily terminal, emphasizing the product. Consequently, the objective of this study consisted of setting up a general process for the formative evaluation of the learning from the evaluative activities carried out by teachers in the classroom. Methodologically, circumscribed to the qualitative paradigm using the ethnographic method, which applied following the stages suggested by Rodriguez, Gil and García (1999): preparatory, field, analytical e informative. To gather the information used participant observation and field as a tool journal to record what the information. The information collected is transcribed, organized, categorized and validated by means of triangulation of sources. As a result obtained a set of underlying activities of formative evaluation in teaching practice of teachers, which integrated into a sequence to form a general process of formative evaluation of learning included in the processes of teaching and learning. The process is useful to perform a conscious and systematic formative evaluation.

Palabras clave: Formative assessment, Research, Formative assessment activities, General process, Elementary school.

*Contacto: evalidpasek@gmail.com
issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 4 de junio de 2016
1ª Evaluación: 21 de junio de 2016
Aceptado: 9 de agosto de 2016

1. Introducción

En todo proceso educativo se pretende formar un individuo según el modelo político de cada país, preparado para que pueda enfrentarse al momento histórico que le corresponde vivir e integrarse en un mundo cambiante, cada vez más complejo y multicultural. En consecuencia, valorar el aprendizaje logrado es uno de los retos más importantes pues se trata de evidenciar la correspondencia de los resultados del proceso educativo con lo preceptuado y esperado. Por eso, la evaluación es concebida como

> *un proceso sistemático, sistémico, participativo y reflexivo que permite emitir una valoración sobre el desarrollo de las potencialidades del y la estudiante, para una toma de decisiones que garantice el logro de los objetivos establecidos en el Currículo Nacional Bolivariano.* (Ministerio del Poder Popular para la Educación, MPPE, 2007, p. 67)

El mismo documento establece tres tipos de evaluación: Inicial y/o diagnóstica, previa a la planificación de enseñanza aprendizaje; la procesual y/o formativa que tiene lugar a lo largo del proceso educativo, y, la final y/o sumativa que tiene la función de valorar los logros alcanzados por los estudiantes al final de un proceso con el fin de promover o certificar conocimientos y habilidades. Así, asume una concepción procesual orientada a la toma de decisiones (Stufflebeam y Shinkfield, 1987).

En el contexto del cambio de paradigma educativo, en Venezuela la evaluación del aprendizaje es cualitativa en primaria (1º a 6ª grados) y está indisolublemente vinculada con el proceso de enseñanza aprendizaje, de tal manera que no hay momentos de evaluación y momentos de enseñanza (Vergara, 2012; Pasek, 2009). Este tipo de evaluación es formativa y responde a una concepción constructivista de enseñanza y aprendizaje que considera el aprender como un proceso en el cual el estudiante va reestructurando su conocimiento a partir de las actividades realizadas.

En el contexto del aula, la evaluación formativa es un proceso que aplican los docentes y estudiantes durante la instrucción y provee información para ajustar tanto la enseñanza como el aprendizaje (Heritage, 2011; López, 2009; TIAL, en Martínez, 2012; Smarter Balanced Assessment Consortium, 2015). Es decir, consiste en un conjunto de actividades orientadas a la identificación de errores, comprender sus causas y tomar decisiones para superarlas desde y en el aula. Las actividades permiten a los docentes recabar y usar la información para una educación integral considerando las necesidades individuales de los niños (Riley-Ayers, 2014).

Sin embargo, en muchas instituciones de educación básica venezolanas la evaluación formativa no se practica de forma sistemática. (OECD, 2005; Pérez, 2005, Villamizar, 2005). Según Pérez (2005) mantiene indicios de carácter tradicional pues, el docente entiende que su papel consiste en la elaboración y aplicación de instrumentos con fines evaluativos (de medición) con base en los cuales establece un juicio valorativo. En otras palabras, todavía no se ha logrado comprender y aplicar un sistema de evaluación vinculado a los planteamientos constructivistas, participativos, donde los docentes evalúen para mejorar su praxis educativa y el proceso de aprendizaje de los educandos.

Aun cuando se acepta que la evaluación formativa es el modo de evaluar idóneo, pues, mejora el proceso de aprendizaje, permite el monitoreo y la realimentación, favorece la identificación de las dificultades (Pereira y Flores, 2016), persiste un enfoque tradicional de evaluación en la escuela venezolana al reducirla a una etapa del proceso de

aprendizaje necesariamente terminal, con énfasis en el producto. Concretamente, en la escuela ámbito del estudio, las investigadoras fueron testigos de reclamos de representantes y estudiantes al manifestar su desacuerdo e inconformidad con los juicios emitidos sobre su comportamiento académico. Los reclamos aumentan al desconocer los motivos del juicio recibido, pues en sus boletines solo aparecen descripciones someras de la apreciación del docente. También en las reuniones de padres se generan debates cuando se señala el bajo nivel académico de una parte importante del estudiantado.

Posiblemente esto se debe a factores como la poca satisfacción de las necesidades de actualización continua que obtienen los docentes de las actividades formativas que imparte el despacho educativo sobre evaluación cualitativa (Alcedo y Chacón, 2010); la confusión que generan los juicios descriptivos (producto de evaluación formativa) y el término de calificación (resultado de evaluación sumativa, cuantitativa), tal como lo develaron las investigaciones realizadas por Pérez (2011) y Román (2011), quienes encontraron la persistencia de una práctica evaluativa alejada del constructivismo, centrada en una abundancia de exámenes, pruebas y otros instrumentos basados mayormente en la medición cuantitativa. Otra posible causa es la conceptualización errada que poseen los docentes sobre la evaluación formativa. En este orden de ideas, Pasek de Pinto y Mejía (2014) realizaron un estudio donde encontraron que los docentes entrevistados entienden la evaluación formativa como realimentación, como función y poseen ciertas visiones erradas; sumado a esto, ninguno la definió como un proceso, es decir, no la perciben sistemática. Igualmente, Mejía (2015) encontró que la concepción que utilizan los docentes en su práctica evaluativa está pensada, primero como realimentación inmediata y, segundo, como actividad con diferentes funciones; muchos la realizan si lo creen necesario.

En vista de lo anteriormente expuesto y partiendo de las actividades de evaluación que realizan los docentes en el aula, esta investigación tuvo como objetivo configurar un proceso general para la evaluación formativa del aprendizaje. Su relevancia radica en que proporciona una secuencia de pasos que permite a los docentes realizar una evaluación formativa de los aprendizajes sistemática, contextualizada e integrada a los procesos de enseñanza aprendizaje. Cabe destacar, que en la revisión bibliográfica realizada encontramos que la mayoría de los autores definen la evaluación formativa como un proceso, sin embargo, son escasos quienes describen los pasos que éste implica. Aunado a ello, el proceso que muchos señalan es el mismo de la evaluación en general, es decir: recoger información, analizarla, emitir juicios y, sobre la base de juicios, tomar decisiones de mejora. Pero, no encontramos un proceso que sistematice específicamente el proceso de la evaluación formativa que se realiza en el aula, considerando unos pasos distintos del proceso de la evaluación en general.

2. Bases teóricas

2.1. Evaluación formativa

Para este estudio se partió de la concepción que plantea el Ministerio del Poder Popular para la Educación (MPPE, 2007, p. 68): la evaluación procesual y/o formativa "se planifica con la finalidad de obtener información de los elementos que configuran el desarrollo del proceso educativo de todos y cada uno de los y las estudiantes,

proporcionando datos para realimentar y reforzar los procesos". Ya que el mismo documento define la evaluación como un proceso, la evaluación formativa debe ser sistemática e implementada por el docente en colaboración con los estudiantes con el fin de obtener la información requerida y pertinente que permita conocer sus avances, dificultades y orientarlos para darle solución durante el desarrollo del proceso de enseñanza aprendizaje. De esta manera, además de orientar hacia el éxito a los alumnos, los docentes pueden ajustar sus programas y mejorar su propia práctica educativa. Cabe destacar que la mayoría de los autores analizados comparten esta concepción sobre la evaluación formativa.

2.2. Proceso de evaluación formativa

En la revisión bibliográfica realizada sobre evaluación formativa encontramos múltiples estudios relacionados con sus características, componentes, funciones, estrategias, actividades sugeridas para ponerla en práctica, relaciones entre la evaluación diagnóstica, formativa y sumativa, entre otros. Pocas investigaciones la definen como una actividad, pues, la mayoría la entiende como un proceso, pero ninguno señala explícitamente unos pasos que puedan constituir un proceso.

A manera de ejemplo, Rosales (2014, p. 5) expresa que por su carácter "La evaluación formativa nos facilita la tarea de identificar problemas, mostrar alternativas, detectar los obstáculos para superarlos, en definitiva, perfeccionar el proceso educativo", característica que por su forma podría asumirse como un proceso de tres pasos. Otro ejemplo es el de Smarter Balanced Assessment Consortium (2015), quienes indican que la evaluación formativa posee cuatro atributos: 1) aclarar el aprendizaje que se quiere lograr, lo que implica establecer metas de aprendizaje y criterios de éxito; 2) obtener evidencia de diversas fuentes en función de las metas y criterios previos; 3) interpretar la evidencia, es decir, dilucidar junto con los estudiantes el significado de la información recabada para determinar dónde ubican su aprendizaje con respecto a las metas y criterios previstos; y por último, 4) actuar sobre la base de la evidencia, o sea, realimentar el proceso y decidir los pasos siguientes con el fin de avanzar en el aprendizaje, tomando en cuenta dificultades, intereses y preferencias de cada estudiante. Desde nuestra perspectiva, ambos procesos son bastante generales, no involucran un monitoreo constante por lo que no ofrecen una realimentación inmediata, indispensable para alcanzar los aprendizajes previstos en cada jornada de trabajo en el aula de clases.

3. Metodología

Esta investigación se realizó en dos etapas que se describen a continuación.

3.1. Primera etapa o de investigación cualitativa

Para conocer las actividades evaluativas que realizan los docentes en el aula, la investigación asumió el paradigma cualitativo. En este paradigma, se estudia la realidad en su contexto natural, tal y como sucede, con la intención de interpretar los fenómenos de acuerdo con los significados que tienen para las personas involucradas, describiendo la rutina y situaciones problemáticas que se les presentan.

Dentro del paradigma cualitativo se asumió la etnografía como método. Ésta, de acuerdo con autores como Goetz y Lecompte (1988), Rodríguez, Gil y García (1999), Gurdián-

Fernández (2007), se caracteriza por ser holística y contextual, sus observaciones son puestas en una perspectiva amplia y asume que el comportamiento de la gente sólo puede ser entendido en su contexto específico. Por consiguiente, un estudio etnográfico busca construir un esquema teórico que recoja y refleje lo más fielmente posible la realidad de la actuación humana, en este caso, el proceso de evaluación formativa que tiene lugar en las clases. En ese orden de ideas, la investigación se realizó siguiendo las fases de la investigación cualitativa que señalan Rodríguez, Gil y García (1999): preparatoria, trabajo de campo; analítica e informativa.

3.1.1. Fase Preparatoria

Incluye dos etapas: la reflexiva y la de diseño. En la primera, sobre la base de la experiencia de las investigadoras, se eligió el tópico de interés y se buscó la información relacionada con el estado del arte del tema. Esto culminó en la interrogante: ¿Cómo es el proceso de evaluación formativa que realizan los docentes en el aula? La incógnita marcó la decisión de optar por una etnografía descriptiva y construir el marco referencial del estudio. Seguidamente, en la etapa de diseño, elaboramos el plan de acción para las subsiguientes fases.

3.1.2. Fase trabajo de campo

Tuvo como finalidad realizar las observaciones y recoger la información necesaria. Comprendió dos etapas: el acceso al campo y la recogida productiva de la información que implican la selección del escenario, la elección de informantes, de técnicas e instrumentos.

- ✓ *Acceso al campo.* En este momento seleccionamos la escuela o escenario, atendiendo, en primer lugar al criterio de pertenencia: una de las investigadoras pertenece al contexto y es docente de la institución, lo que facilita el acceso a la información; el segundo criterio fue el tamaño de la escuela, es la más grande del municipio escolar y atiende desde educación inicial o preescolar hasta sexto grado de primaria. Durante la primera visita, conversamos con los 30 docentes de la institución, les informamos del propósito y les propusimos ser partícipes de la investigación permitiendo la observación de sus clases. En un segundo encuentro hablamos con los 10 docentes que se mostraron más receptivos; de ellos, 8 aceptaron voluntariamente formar parte del estudio, sólo uno es de sexo masculino. Todos poseen licenciatura en educación integral, muchos han participado en talleres de formación relacionados con escuelas bolivarianas y los proyectos como estrategia de enseñanza aprendizaje. Ninguno manifestó haber asistido a talleres sobre evaluación. En educación primaria un mismo docente imparte todas las materias del grado. Otras características se muestran en la Tabla 1.

- ✓ *La recogida productiva de información.* Se eligió la técnica de la observación participante para recabar la información. Ésta, siguiendo a Angrosino (2012), es entendida como un procedimiento interactivo de recogida de información que involucra al observador en los acontecimientos o fenómenos que está observando, permitiendo observar sin restricciones todas las clases. Como instrumento para registrar la información se usó el diario de campo. Por consiguiente, se observó a los 8 docentes en sus aulas durante las clases de enero

a marzo de 2014, recogiendo en forma de una narración minuciosa las experiencias vividas y los hechos observados. Es importante señalar que durante la observación sólo se plasmó estrictamente lo observado, sin realizar comentarios ni análisis subjetivos con el fin de conservar el rigor y la objetividad que exige la investigación.

Tabla 1. Características de los docentes observados

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8
Género	Fem.	Fem.	Fem.	Fem.	Fem.	Fem.	Fem.	Masc.
Años de docencia	12	10	5	3	3	6	5	5
Grado y sección que atiende	2°"A"	2°"B"	4°"A"	4° "B"	3° "A"	5° "A"	1°"A"	3°"B"
Postgrado	Maestría en Educ.	Maestría en Educ.	No	No	Espec en docencia	No	No	No

Fuente: Pasek de Pinto y Mejía (2016), con datos suministrados por los docentes.

3.1.3. Análisis de datos

El análisis de datos lo realizamos mediante un proceso dinámico y creativo para descubrir y estructurar las relaciones subyacentes en los hechos observados. Así, establecimos una serie de tareas como lectura y revisión de la información, categorización, triangulación y obtención de hallazgos, conclusiones y reflexiones (Angrosino, 2012; Rodríguez, Gil y García, 1999). La categorización comenzó con una descripción de lo recabado en las notas, luego la clasificamos, identificando categorías o temas. Cada categoría soporta un significado o tipo de significados. Para ello, asignamos un concepto a las actividades evaluativas que realizan los docentes en el aula obteniendo las categorías resultantes, las que describimos desde lo observado.

Para dar validez a la información obtenida y a las categorías construidas acudimos a dos actividades: en primer lugar, la devolvimos a los informantes en una entrevista, es decir, las categorías conformadas las discutimos con cada docente para comprobar que reflejan con fidelidad su modo de actuar. En segundo lugar, realizamos la triangulación de la información contrastándola con varias fuentes como los mismos docentes, teóricos e investigadores de la evaluación y la interpretación de las autoras. El objetivo de la triangulación fue, por un lado, la validación y por otro, contar con una multiplicidad de perspectivas; lo que incrementó la validez de los resultados al depurar las deficiencias intrínsecas de un solo método de recogida de datos y favoreció el control del sesgo personal del investigador.

3.1.4. Fase informativa

En esta fase integramos la fundamentación teórica y empírica que apoya el trabajo y presentamos los resultados obtenidos.

3.2. Segunda etapa o de construcción teórica

En esta etapa sistematizamos la información, la interpretamos; integramos y secuenciamos de una manera lógica ciertos conocimientos que eran, hasta el momento, imprecisos, inconexos o intuitivos y conformamos un proceso general para la evaluación formativa desde el hacer pedagógico de los docentes y los teóricos sobre el tema.

4. Resultados

En este apartado presentamos los resultados en dos partes. En la primera se muestran las actividades de evaluación formativa subyacentes en la práctica pedagógica de los docentes como producto de la observación realizada de las clases que ejecutan los docentes. En la segunda, ofrecemos el proceso inferido y reconstruido de su actuación evaluativa considerando los resultados anteriores.

4.1. Actividades de evaluación formativa subyacentes en la práctica pedagógica

Para darle la claridad y objetividad requeridas, las actividades que realizan los docentes las organizamos en forma secuencial y se presentan en la tabla 2.

Tabla 2. Actividades de evaluación formativa que realizan los docentes observados

ACTIVIDAD	F	DESCRIPCIÓN	VIÑETA
Exploración	7	3 docentes muestran interés en el tema de la clase anterior 3 se centran en el tema que debían desarrollar durante las clases observadas 1 preguntó por los conocimientos previos o necesarios para desarrollar el contenido de cada día.	[O a Doc1]: La inicia su clase preguntando: "¿Qué temas desarrollamos ayer?" [O]: Una vez que la docente retoma la clase anterior y explora lo aprendido por los estudiantes, inicia la clase.
Monitoreo	8	Todos los docentes realizan el monitoreo recorriendo el aula y deteniéndose en cada estudiante o grupo. Identifican fallas, errores e indagan por las causas. Comunican las fallas e indican cómo superarlas. Ofrecen orientación oportuna para corregir.	[O a Doc4]: La docente realiza un dictado sobre el sistema solar; a medida que lo realiza va recorriendo cada uno de los puestos de los niños para verificar que estén escribiendo correctamente las palabras y les corrige. [O a Doc8]: El docente se sentó con el grupo para explicarles.
Resalta Logros	3	Lo realizan individualmente y grupalmente Felicita a todos y manifiesta la importancia de cumplir las tareas. Felicita a la niña por el avance significativo en la lectura. Felicita a todos y expresa satisfacción por sus logros.	[O a Doc3]: Al finalizar la tarea la Doc3 manifiesta su satisfacción ante el resultado obtenido durante el desarrollo de la actividad: "*Excelente trabajo, los felicito, me siento muy contenta porque noto que han entendido por el mapa conceptual que elaboraron*"
Controla	3	El control lo realizaron 1 docente de primer grado y 2 de segundo. Cotejan el progreso utilizando los criterios establecidos. Registran la información en su libreta. Las tres se concentraron en el proceso de lectura y escritura.	[O a Doc7]: La docente selecciona ocho alumnos para tomarle la lectura diaria. Los va llamando uno a uno, los orienta en cada una de las oraciones que leen, luego va registrando en el cuaderno de control de lectura lo observado en cada estudiante.

Fuente: Pasek de Pinto y Mejía (2016) con información de las observaciones de clases.

Tabla 2. Actividades de evaluación formativa que realizan los docentes observados (Continuación)

Promueve la auto-evaluación	3	Las tres docentes que la propician asignan tareas de diferentes materias en el aula, estableciendo un tiempo prudencial para su ejecución. Finalizado éste, escriben los resultados en el pizarrón y piden a los niños que cotejen y corrijan su propia ejecución.	[O a Doc1]: La docente le pide al niño la libreta y le señala que observe lo escrito en el pizarrón y cómo lo escribió en su libreta; el niño se da cuenta de su error. Doc1: *"borra y vuelve a escribirlo"*. N_8: *"Tiene razón, voy a acomodarlo maestra, es que estaba compitiendo con Rita para ver quien escribía más rápido"*.
Promueve la co-evaluación	2	La docente (Doc5) asignó un trabajo grupal. Al finalizar la actividad llamó a uno por equipo para que en voz alta leyera su producción. Luego preguntó a los demás equipos su opinión. La maestra (Doc2) asignó exposiciones grupales. Al terminar cada grupo solicitó su opinión a los otros grupos.	[O a Doc2]: Los niños de la Doc2 realizan exposiciones este día. Luego que cada grupo expone, la maestra pregunta al grupo oyente: *"¿Cómo expusieron sus compañeros?"*
Ofrece realimen-tación	3	La mayoría (5) de los docentes termina la clase asignando alguna actividad para la casa y despidiéndose 3 docentes realizaron cierres de clase para fortalecer lo aprendido mediante recuentos orales de las actividades, resúmenes escritos de los temas tratados en clase, definiciones propias de conceptos tratados en clase. La doc4 llega a la transferencia de conocimiento cuando pidió resolver problemas de matemática involucrando el ciclo hidrológico.	[O a Doc6]: Para culminar la jornada, la docente proporciona a cada estudiante una hoja para que realicen un resumen de los temas desarrollados durante la mañana, con las siguientes instrucciones: *"al terminar de realizar esta actividad cada uno la va a leer en voz alta y luego la guarda en su portafolio."*

Fuente: Pasek de Pinto y Mejía (2016) con información de las observaciones de clases.

4.1.1. Análisis y discusión

En la tabla 2 se puede observar que la exploración la realizan 7 (87,5%) docentes; todos los docentes (8, o sea, 100%) monitorean a los estudiantes durante sus actividades de aprendizaje en el aula; 3 de ellos (37,5%) resaltan sus logros, llevan un control de los avances, promueven la autoevaluación de los estudiantes y ofrecen realimentación como cierre de la clase; 2 (25%) promueven la coevaluación.

Con respecto a la evaluación formativa, el estudio permitió hallar evidencias de que los docentes realizan la evaluación formativa de manera intuitiva, como parte de la clase y sin notar su carácter evaluativo al ofrecer una realimentación inmediata a los estudiantes. Así, encontramos que al inicio de la clase, la mayoría **e**xplora el conocimiento que poseen los estudiantes como base del aprendizaje subsiguiente. Esto se corresponde tanto con los principios del constructivismo como con el Diseño Curricular Bolivariano (MPPE, 2007, p. 70), donde señala que explorar es una función de la

evaluación que "permite obtener evidencias sobre las experiencias de aprendizaje del estudiante, sus alcances en relación con los objetivos educativos; vinculadas al contexto donde se producen".

Todos los docentes monitorean las actividades de los alumnos en el aula, es decir, hacen un seguimiento al desarrollo de las actividades de aprendizaje, lo que les permite obtener información valiosa sobre el logro de los objetivos previstos en el proyecto que se esté ejecutando (Aguilar y Ander-Egg, 1994). Asimismo, todos lo realizan recorriendo las mesas de trabajo, deteniéndose en los estudiantes y/o en los grupos. Durante este recorrido pueden identificar fallas y errores, las comunican a los estudiantes y ofrecen orientación oportuna para su corrección. Sin embargo, sólo una docente preguntó por las causas de las fallas y errores, evidenciando una preocupación por la mejora del aprendizaje, lo que enfatiza una concepción formativa de la evaluación, objeto de reflexión para Riley-Ayers (2014), Pereira y Flores (2016), Vergara (2012) y conforme con los planteamientos del MPPE (2007).

La orientación oportuna se entiende como una actividad fundamental en el proceso evaluativo, pero sólo tres docentes la realizan, destacando su función formativa (Heritage, 2011; Smarter Balanced Assessment Consortium, 2015): una revisa los cuadernos de los niños llamándolos uno por uno; otra, durante el recorrido, observa y corrige inmediatamente los errores; el tercero, se sienta con los estudiantes y les explica.

En el transcurso de la clase y durante las diferentes actividades que realizan los niños en el aula, solamente tres de los docentes se detienen a destacar los logros que van alcanzando los estudiantes, tanto de manera individual como grupal. Las felicitaciones y expresiones de satisfacción de las maestras son importantes y necesarias motivaciones para seguir aprendiendo; se trata de hacerle saber al estudiante que va en la dirección correcta para lograr los objetivos previstos (Smarter Balanced Assessment Consortium, 2015).

Llevar el control de los aprendizajes es importante para emitir juicios informados. Este control se puede ver como una verificación de resultados (Aguilar y Ander-Egg, 1994), o bien como el registro de los aprendizajes que van logrando a medida que avanza el desarrollo de los proyectos (Riley-Ayers, 2014; Smarter Balanced Assessment Consortium, 2015;). Por eso, los docentes deben supervisar constantemente el avance de los alumnos. Sin embargo, durante la investigación se detectó que el registro sistemático lo llevaron sólo 3 de las docentes. Se puede inferir que este hecho es una posible causa de la insatisfacción de representantes y estudiantes al recibir en la boleta la somera información que describe su actuación.

Promover la autoevaluación es un aspecto primordial para la formación puesto que involucra un proceso de reflexión (MPPE, 2007; Smarter Balanced Assessment Consortium, 2015). No obstante, solamente lo realizan 3 de los docentes observados, por lo que se puede inferir que no se está formando el ciudadano reflexivo y crítico que preceptúa la Constitución de la República Bolivariana de Venezuela (1999). Lo mismo ocurre con promover la coevaluación, la cual está concebida como una forma de participar en la evaluación, es decir, los estudiantes aprenden a ser ciudadanos participativos mediante la valoración que realizan sobre la actuación de los compañeros. Pero solamente dos de los docentes observados propician actividades para la

coevaluación, en consecuencia, los alumnos no están aprendiendo a ser participativos ni a valorar el trabajo del otro.

En lo que se refiere a la realimentación, según el currículo nacional bolivariano (MPPE, 2007), ésta permite tomar decisiones para reorientar y fortalecer las acciones educativas ejecutadas sobre la base de los avances y logros obtenidos. Asimismo, promueve la reflexión sobre sus actos y consecuencias, les permite vislumbrar las metas y objetivos de la planificación; por ello, es importante que los docentes realicen un cierre de clase que fortalezca lo aprendido. Sin embargo, sólo 3 docentes la realizaron solicitando a los niños un recuento oral o un resumen escrito de los temas trabajados; uno llegó a la transferencia. Se puede inferir que no se cierra el ciclo del aprendizaje, no se afianzan los conocimientos, lo que puede incidir en el desempeño académico de los estudiantes, objeto de reclamo de los padres y representantes.

En resumen, es importante señalar que una sola docente realizó todas las actividades descritas, 3 docentes realizaron actividades de evaluación formativa y ninguno de los docentes ve tales actividades como evaluativas, pues las asumen como actividades de enseñanza aprendizaje. En consecuencia, configuramos un proceso para la evaluación formativa contextualizado, utilizando para ello las actividades ordenadas en una secuencia lógica y aplicable en el aula.

4.2. Construcción de un proceso general para realizar una evaluación formativa en el aula

El proceso general para la evaluación formativa generado está estrechamente vinculado con los procesos de enseñanza aprendizaje, se contextualiza en el ámbito del aula y se centra en el aprendizaje del estudiante, como se puede ver en la figura 1.

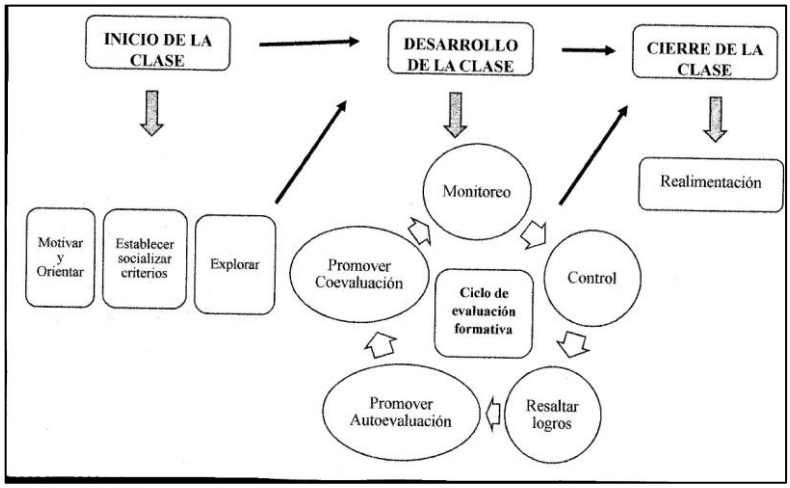


Figura 1. Proceso general de evaluación formativa
Fuente: Elaborado por las autoras.

En la figura 1 se observa que el proceso de la evaluación formativa se fundamenta en los tres pasos del proceso de enseñanza: inicio, desarrollo y cierre de la clase.

4.2.1. El inicio de toda clase

Está dirigido a despertar el interés de los estudiantes e implicarlos en el tema objeto de estudio. Por eso, esta etapa de la enseñanza contiene el motivar y orientar, el establecer y acordar los criterios de evaluación y la exploración de saberes y conocimientos; incluye plantear el objetivo o los objetivos de aprendizaje.

- ✓ *Motivar y orientar*: En este paso es importante presentar a los estudiantes los contenidos por aprender, los objetivos a lograr, así como las actividades que se van a realizar para ello. Esto permitirá que los estudiantes conozcan con anticipación el objetivo de aprendizaje que se persigue, el tipo de actividades que planificó el docente, el tiempo destinado para su realización, la importancia que tiene cada actividad para el logro de los aprendizajes esperados.
- ✓ *Establecer y socializar criterios de evaluación*: Toda evaluación requiere unos criterios que le sirvan como base para comparar el avance de los estudiantes, por eso es importante definir y compartir con sus estudiantes los que utilizará para valorar sus logros. Sólo así tendrán claro lo que se espera de ellos al terminar cada actividad y cada jornada; a la vez, participarán activa y reflexivamente en su proceso de aprendizaje, evitando reclamos posteriores. Es importante señalar que los criterios de evaluación se pueden originar en los contenidos, en los objetivos de aprendizaje y/o en las competencias, capacidades y actitudes que debe desarrollar el estudiante en determinada área curricular.
- ✓ *Exploración:* Explorar es una función de la evaluación que permite obtener evidencias sobre las experiencias de aprendizaje del estudiante, sus conocimientos previos sobre un tema. La exploración puede tomar formas diferentes: puede centrarse en el tema de la clase anterior, indagar por la preparación para el tema que va a desarrollar, o bien, preguntar por los conocimientos previos o necesarios para desarrollar el contenido del día.

En todo caso, es útil para orientar su trabajo teniendo presente hacia dónde debe dirigirse y cuál es la mejor forma para llegar hasta allí ya que permite ubicar el nivel de aprendizaje del estudiante con respecto al formulado en los objetivos. Además, la exploración es necesaria para elegir las estrategias de aprendizaje, pues, el docente debe establecer puentes entre lo que conoce un estudiante y lo que puede y/o debe aprender, fortaleciendo y apoyando el proceso de construcción del conocimiento.

4.2.2. El desarrollo de las clases

En el proceso de enseñanza aprendizaje, es el tiempo planificado por el docente para propiciar el aprendizaje de los contenidos mediante el uso de diversas estrategias, es decir, es el tiempo dedicado a la tarea, a concretar las actividades de aprendizaje. Ya sea que los estudiantes estén organizados en grupos o trabajen individualmente, es el período cuando el docente observa el desenvolvimiento de los alumnos y los orienta hacia el logro de metas y objetivos. Es el momento de la evaluación formativa del aprendizaje y se puede afirmar que cumple un ciclo que se inicia con cada nuevo estudiante o grupo que observa el docente en su recorrido por el aula. Sus pasos son:

- ✓ *Monitoreo.* Consiste en hacer un seguimiento constante a los avances en el dominio de conocimientos o desarrollo de habilidades convenidas en criterios

previamente acordados. Su propósito consiste en observar y analizar las acciones durante el proceso de aprendizaje del estudiante y, de ser necesario, ofrecer nuevas opciones. Está ligado al tiempo y los esfuerzos requeridos para culminar una determinada actividad y sus ajustes. Por lo general esta actividad la realizan los docentes desplazándose por el aula y deteniéndose en cada estudiante, mesa o grupo de trabajo con la finalidad de observar y estimar su desempeño.

Durante el recorrido por las mesas de trabajo, al observar las realizaciones de los estudiantes, los docentes pueden identificar fallas y errores e indagar por las causas, lo que les permite comprender mejor el proceso de cada alumno en tanto individuo único que es y poder ayudarlo. En ese sentido, es importante comunicar las fallas o errores e indicarles cómo superarlas para que tomen conciencia y realicen las correcciones pertinentes. Es decir, el docente debe ofrecer la orientación oportuna para que corrijan. Esta orientación consiste en atender la situación o debilidad detectada de manera inmediata y adecuada, indicando el curso correcto de la actividad. Es una manera de realimentación del aprendizaje que forma parte del desarrollo de la clase y debe ser inmediata, continua y relevante (López, 2010, citado por Osorio y López, 2014). Hay diferentes maneras de orientar en el aula: mediante la revisión de los cuadernos de los niños, convocar uno por uno, acudir en su ayuda cuando llaman, responder a las dudas en el momento que las presenten.

- ✓ *Control:* El control consiste en verificar los resultados de una actividad con miras al logro de los objetivos previstos. Implica cotejar y registrar los progresos respecto de los criterios establecidos para cada caso. Por eso, llevar el registro de los aprendizajes alcanzados por los niños es de suma importancia e involucra una supervisión constante del proceso de aprendizaje. Para verificar los logros se requieren los criterios establecidos al inicio de la clase y socializados con ellos, así como los distintos instrumentos de evaluación, entre otros, registro diario, lista de cotejo, escala de estimación. Cabe destacar la importancia de llevar un control sistemático considerando que ese registro servirá de insumo para emitir los juicios descriptivos y valorativos de la actuación del estudiante y es útil para aclarar dudas a estudiantes, padres y representantes.
- ✓ *Resaltar logros.* Se trata de enfatizar los resultados positivos de los estudiantes en el desarrollo de su aprendizaje. En el aula, durante la verificación o control de una actividad nos encontramos con potencialidades, fortalezas, avances y, por supuesto, debilidades. Resaltar logros consiste en manifestar de forma oral o escrita satisfacción por la actividad realizada, por el cumplimiento del objetivo propuesto, por el desempeño del estudiante en una responsabilidad asignada; es felicitarlos por la labor cumplida, entre otros. Su importancia reside en que el estudiante entiende que está bien encaminado hacia la meta.

  Sin embargo, es importante que además de destacar los logros, expresemos muy sutilmente sugerencias para superar las debilidades halladas ya que representan palabras de aliento para continuar esforzándose. En ambos casos nos valemos del uso de los indicadores que establecimos para los criterios de evaluación. Los indicadores son las pistas observables del desempeño de los niños y proporcionan información cualitativa o cuantitativa sobre el rasgo que se valora

pues describen de manera real el desarrollo de las capacidades y actitudes. Su relevancia radica en el hecho de que pasan a conformar los juicios descriptivos que emitimos y asentamos en las boletas de información.

- ✓ *Promover la autoevaluación.* La autoevaluación "es el proceso de reflexión que realiza cada uno de los participantes responsables del proceso de aprendizaje". (MPPE, 2007, p. 68). Constituye un objetivo de la evaluación formativa y se basa en que existan criterios de evaluación claros y explícitos para que los estudiantes puedan pensar, aplicar y reflexionar en el contexto de su propio trabajo. Así mismo, es una forma de participar en la evaluación que favorece el desarrollo de procesos de autorregulación, capacidad muy importante en el funcionamiento individual y social de las personas.

  Con la autoevaluación se pretende que el estudiante mismo tome conciencia de sus fortalezas y debilidades, aciertos y errores. En ese sentido, se trata de orientar su atención hacia los aciertos y errores sin respuestas elaboradas. Tal vez sería adecuado enseñar a los estudiantes el proceso de reflexión que plantea Smyth (1991, p. 279), consistente en aplicar los cuatro pasos siguientes: "1. Descripción: ¿Qué es lo que hago?; 2. Inspiración: ¿Cuál es el sentido de la enseñanza que imparto?; 3. Confrontación: ¿Cómo llegué a ser de esta forma?; y 4. Reconstrucción: ¿Cómo podría hacer las cosas de otra manera?", adaptándolos a cada actividad sobre la que se pretenda reflexionar para autoevaluarse. Este proceso permite recrear la actuación, errada o acertada; pensar hacia dónde lo lleva la acción; analizar por qué la realizó así (acertada o errada) y, finalmente, buscar soluciones que le permitan hacerlo mejor. Con ello tendríamos los estudiantes reflexivos y críticos que preceptúa el Ministerio del Poder Popular para la Educación (2007).

- ✓ *Promueve la coevaluación.* La coevaluación consiste en la evaluación mutua, conjunta de una actividad o un trabajo realizado entre varias personas como, por ejemplo, de los estudiantes entre sí. La coevaluación se puede realizar por grupos pequeños cuando el trabajo en el aula es cooperativo; en su recorrido, el docente, al pasar por los grupos puede preguntar a cada uno cómo ha sido el desempeño y cuál ha sido el aporte de los otros al logro del objetivo y culminación de la actividad. En otros casos se realiza con la sección completa, por ejemplo, cuando docentes y estudiantes han planificado exposiciones, entonces, el docente le puede preguntar al resto de la sección sobre la exposición del grupo.

Cabe señalar que concentrar la atención sobre la actuación y/o tareas realizadas por los compañeros de la clase permite descubrir las propias dificultades, o también, fortalezas. Sin embargo, es importante que la coevaluación se realice considerando los criterios establecidos al inicio de la clase, que se refieran al desempeño y no a la persona y que se eviten calificativos ofensivos o discriminativos.

4.2.3. El final de la clase

Representa los últimos minutos, requiere un cierre y no es suficiente con asignar las tareas para la próxima clase y despedirse. Es el momento de la realimentación y no forma parte del ciclo de la evaluación formativa iniciado con el monitoreo.

- ✓ *Realimentación.* Consiste en retomar los contenidos para afianzarlos, comunicar a los estudiantes los errores más frecuentes, los obstáculos al conocimiento que ha detectado, brindar orientaciones claras sobre cómo mejorar en sus desarrollos, logros y desempeños. Es el momento de aclarar las dudas que aún persisten, de generar conclusiones. La realimentación permite que tanto el docente como los estudiantes tomen conciencia del logro del objetivo de la clase, de sus progresos, de sus nuevos aprendizajes; también sirve de base a la toma de decisiones para reorientar y/o fortalecer las acciones educativas ejecutadas.

Considerando que el proceso de aprendizaje pasa por tres momentos: adquisición o construcción, repaso y transferencia de conocimientos, la realimentación es el tiempo para detectar el aprendizaje realmente logrado sobre la base de la transferencia de conocimientos que realiza el estudiante a otras áreas y/o a otros contextos. Entre otras dinámicas adecuadas para este momento, se puede formular preguntas a la sección sobre la experiencia de lo vivido en la clase, pedir a los estudiantes que escriban una síntesis y la compartan, leer las conclusiones del trabajo realizado en el grupo, pedirles que piensen en las posibles aplicaciones de lo aprendido en otras materias, áreas, o a su vida cotidiana. Además, es importante aprovechar el momento para enseñarles a que se realimenten unos a otros.

5. Conclusiones

La evaluación formativa es entendida por los investigadores como el proceso valorativo idóneo para mejorar la educación en general y los procesos de enseñanza aprendizaje en particular. Sin embargo, su aplicación intuitiva y asistemática se aleja de los planteamientos de la evaluación formativa constructivista y participativa establecida. Por tal motivo, se pretendió hallar evidencia de que en la práctica evaluativa del docente en el aula subyace un proceso de evaluación formativa que es posible explicitar y configurar en un diagrama que muestre su sistematicidad.

Durante la experiencia de investigación se encontró que, desde el punto de vista teórico, la evaluación formativa es definida como un proceso estrechamente vinculado con los procesos de enseñanza aprendizaje. Por otra parte, desde la perspectiva de la práctica, pocos docentes realizan actividades de evaluación formativa y quienes lo hacen, las asumen como actividades de enseñanza aprendizaje que son parte de la clase y no constituyen pasos de un proceso. Esto posiblemente se debe a que el monitoreo y la orientación inmediata al estudiante los realizan durante la ejecución de las tareas de aprendizaje diseñadas y orientadas por el docente para/en el aula, por lo que tampoco toman nota de la información que generan.

Lo anterior trae como consecuencias que, si bien, todos los docentes hacen un seguimiento al desarrollo de las actividades de aprendizaje, son pocos los que registran sistemáticamente la información y carecen de ella cuando deben formular juicios de la actuación del estudiante. Si los registros son incompletos, sus juicios descriptivos pueden estar mal fundamentados y ser objeto de reclamos por parte de estudiantes y padres o representantes. Igualmente puede afectar las decisiones que tomen respecto de nuevas estrategias, actividades, en los futuros pasos para alcanzar los criterios de éxito establecidos.

Tanto la autoevaluación como la coevaluación son modos de participar en el proceso educativo. No obstante, la escasa práctica de dichas formas de evaluación limita el aprendizaje de la participación de los estudiantes, por lo que es posible afirmar que, desde una evaluación formativa vista así, no se está formando el ciudadano participativo que tiene por meta la educación.

Un aspecto álgido lo representa la realimentación o cierre de la clase. Se entiende que este momento no es sólo para asignar tareas y despedirse; por el contrario, su esencia radica en cerrar el ciclo del aprendizaje que pasa por la adquisición o construcción de nuevos conocimientos, el repaso de lo aprendido y, finalmente, llegar a la transferencia de lo aprendido a otras materias, otros contextos y, aún, a la vida cotidiana. Si esta transferencia no se realiza, como se observó, no se afianzan los conocimientos y no es posible hablar de aprendizaje logrado.

En síntesis, dado que en la práctica se encontró escasa aplicación de la evaluación formativa aunado a que es poco sistemática, se configuró como resultado de investigación un proceso general para realizar una evaluación formativa del aprendizaje consciente y metódica. Se exhorta a los docentes a probar su utilidad en el contexto de las clases para obtener más y mejor información sobre el progreso de cada estudiante. Si no aplicamos una evaluación que realmente forme, lograr una escuela/aula formadora de personas participativas, creativas, integrales quedará fuera de alcance en nuestro futuro.

Referencias

Aguilar, M. y Ander-Egg, E. (1994). *Evaluación de servicios y programas sociales*. Buenos Aires: Lumen.

Alcedo, Y. y Chacón, C. (2010). El rol del docente en la evaluación cualitativa de los aprendizajes en inglés. *Revista Evaluación e Investigación*, *5*(1), 55-72.

Angrosino, M. (2012). *Etnografía y observación participante en investigación cualitativa*. Madrid: Morata.

Goetz, J. P. y LeCompte, M. D. (1988). *Etnografía y diseño cualitativo en investigación educativa*. Madrid: Morata.

Gurdián-Fernández, A. (2007). *El paradigma cualitativo en la investigación socio-educativa*. San José de Costa Rica: Coordinación Educativa y Cultural Centroamericana (CECC) y Agencia Española de Cooperación Internacional (AECI).

Heritage, M. (2011). Formative assessment: An enabler of learning. *Better: Evidence-Based Education*, *6*, 18-19.

López, V. (2009). *Evaluación formativa y compartida en educación superior*. Madrid: Narcea.

Martínez Rizo, F. (2012). La evaluación formativa del aprendizaje en el aula en la bibliografía en inglés y francés. Revisión de literatura. *Revista Mexicana de Investigación Educativa*, *17*(54), 849-875.

Mejía, M. T. (2015). *Procedimientos para la evaluación formativa bajo la filosofía constructivista del aprendizaje en educación básica* (Tesis doctoral). Universidad Fermín Toro, Venezuela.

OECD. (2005). *Formative assessment improving learning in secondary classrooms: Improving learning in secondary classrooms*. París: OECD Publishing.

Osorio, K. y López, A. (2014). La retroalimentación formativa en el proceso de enseñanza-aprendizaje de estudiantes en edad preescolar. *Revista Iberoamericana de Evaluación Educativa*, *7*(1), 13-30.

Pasek de Pinto, E. (2009). Reflexiones en torno a la evaluación cualitativa en el aula. *Academia*, *8*(16), 2-12.

Pasek de Pinto, E. y Mejía, M. T. (2014). ¿Cómo entienden la evaluación formativa los docentes? *Academia*, *7*(31), 53-69.

Pereira, D. y Flores, M. (2016). Conceptions and practices of assessment in higher education: A study of Portuguese university teachers. *Revista Iberoamericana de Evaluación Educativa*, *9*(1), *9-29*. doi:10.15366/riee2016.9.1.001

Pérez, E. (2005). Enseñanza y evaluación: Lo uno y lo diverso. *Educere*, *9*(31), 473-479.

Pérez, M. (2011). *Plan de formación docente sobre la evaluación como acción pedagógica en Educación Inicial* (Trabajo de Grado de Maestría). Universidad Valle del Momboy, Venezuela.

Riley-Ayers, S. (2014). *Formative assessment: Guidance for early childhood policymakers*. New Brunswick, NJ: Center on Enhancing Early Learning Outcomes.

República Bolivariana de Venezuela. (1999). *Constitución de la República Bolivariana de Venezuela*. Gaceta Oficial, núm. 36.860, de 30 de diciembre de 1999. Recuperado de http://www.uc.edu.ve/archivos/constitucion.PDF

Ministerio del Poder Popular para la Educación. (2007). *Currículo Nacional Bolivariano. Diseño Curricular del Sistema Educativo Bolivariano*. Recuperado de http://www.me.gob.ve/media/contenidos/2007/d_905_67.pdf

Rodríguez, A., Gil, J. y García, E. (1999). *Metodología de la investigación cualitativa*. Málaga: Aljibe.

Román, F. (2011). *Plan de formación docente sobre el uso adecuado de instrumentos de evaluación cualitativa en Educación Básica* (Trabajo de Grado de Maestría). Universidad Valle del Momboy, Venezuela.

Rosales, M. (noviembre, 2014). *Proceso evaluativo: Evaluación sumativa, evaluación formativa y assesment. Su impacto en la educación actual*. Trabajo presentado en el Congreso Iberoamericano de Ciencia, Tecnología, Innovación y Educación, Buenos Aires. Recuperado de www.oei.es/congreso2014/memoriactei/662

Smarter Balanced Assessment Consortium. (2015). *The formative assessment process*. Recuperado de https://www.smarterbalanced.org/wp-content/uploads/2015/09/Formative-Assessment-Process.pdf

Smyth, J. (1991). Una pedagogía crítica de la práctica en el aula. *Revista de Educación*, *294*, 275-300.

Stufflebeam, D. y Schinkfield, A. (1987) *Evaluación sistemática. Guía teórica y práctica*. Barcelona: Paidós.

Vergara, C. (2012). Análisis de las concepciones de evaluación del aprendizaje de docentes destacados de educación básica. *Revista Iberoamericana de Evaluación Educativa*, *5*(3), 251-274.

Villamizar, J. (2005). Los procesos en la evaluación educativa. *Educere*, *9*(31), 541-544.

Breve CV de los autores

Eva Pasek de Pinto

Licenciada en Educación, mención Ciencias Biológicas; Especialista en Metodología de la investigación y Magíster en Planificación y Administración de la Educación Superior; Doctora en Ciencias de la Educación. Actualmente jubilada por la Universidad Nacional Experimental Simón Rodríguez. Como investigadora en las áreas de evaluación, ambiente, y, el conocimiento y sus obstáculos, es miembro de las Líneas de Investigación "Investigadores en Acción Social (IAS)" y "Fortalecimiento de la Educación Inicial (LinFEI)". Posee publicaciones en revistas nacionales e internacionales. De su experiencia docente e investigadora publicó un libro: *Didácticos. Ideas. Investigación. Educación* como autora-editora. Es investigadora reconocida Nivel C del Programa de Estímulo a la Investigación e Innovación (PEII). ORCID ID: 0000-0002-6471-2467. Email: evalidpasek@gmail.com

María Teresa Mejía

Licenciada en Educación Integral, MsC. Administración de la Educación Básica, Doctora en Ciencias de la Educación. Docente de aula en educación básica. Profesora invitada de Post-Grado en las universidades: Universidad Valle del Momboy, Rafael María Baralt y Universidad Pedagógica Experimental Libertador, extensión Rubio. Posee publicaciones en revistas nacionales. Email: mariateresa_mejia@hotmail.com

Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 195-210.
https://doi.org/10.15366/riee2017.10.1.010

La Evaluación del Conocimiento Metalingüístico en Niños del Último Ciclo de la Educación Infantil Peruana

Assessment of Metalinguistic Knowledge in Children of Last Cycle of Peruvian Preschool Education

Liz Ysla Almonacid [1]*
Vicenta Ávila Clemente [2]

[1] Instituto de Educación Superior Pedagógico Privado Calidad en Redes de Aprendizaje – CREA
[2] ERI-Lectura, Universitat de València

El conocimiento metalingüístico es la capacidad para reconocer la naturaleza, formas y funciones del lenguaje escrito. Implica la toma de conciencia de lo impreso. Aunque existen mayores referentes sobre su evaluación en lengua inglesa, crece el interés por investigar qué sucede en niños de habla española. El propósito del presente trabajo es evaluar las tareas del conocimiento metalingüístico en niños de 5 años con el empleo de una de las tareas que contiene la Batería de Inicio a la Lectura-BIL (dirigido a niños entre 3 y 6 años). Participaron 90 niños del aula de 5 años. Fueron distribuidos en tres grupos de edad propuestos por la BIL. Se evaluaron tres tareas: reconocimiento de palabras, reconocimiento de frases y funciones de la lectura. Los resultados confirman que se trata de una habilidad presente en los niños del aula de 5 años. Se evidencia su condición evolutiva, al encontrarse diferencias significativas en el reconocimiento de palabras y conocimiento de las funciones de la lectura entre los tres grupos de edad. Asimismo, ha sido posible el empleo de un instrumento diseñado en idioma español.

Palabras clave: Conocimiento metalingüístico, Competencia lectora, Habilidades prelectoras, Evaluación del lenguaje infantil, Conciencia fonológica

Metalinguistic knowledge is defined as the ability to recognise components of written language as well their nature, form and function. It involves print awareness. Although there is more evidence in English language about metalinguistic knowledge assessment, there is a growing interest to investigate how it develops in Spanish-speaking children. The purpose for this paper is to assess the tasks of this predictor in five-years-old children with the use of one of the tasks contained in *Batería de Inicio a la Lectura –BIL*; an instrument aimed at children between 3 to 6 years. A total of 90 children were evaluated. All of them attend to five-years-old classroom and were divided into three age groups according to the test. There were three tasks evaluated: recognition of words and phrases and reading functions. Findings confirm that knowledge metalinguistic can be observed in this age group. In addition, their evolutionary condition has been confirmed. There were significant differences in word recognition and reading functions between age groups. It was possible to have applied an instrument designed to Spanish-speaking children.

Keywords: Metalinguistic knowledge, Reading literacy, Prereading skills, Child language assessment, Phonological awareness.

*Contacto: liz.ysla@gmail.com
issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 21 de junio de 2016
1ª Evaluación: 12 de agosto de 2016
Aceptado: 24 de octubre de 2016

1. Introducción

Reconocidas también como predictores o facilitadores de la lectura (Sellés y Martínez 2008), las habilidades prelectoras favorecen la alfabetización inicial y el desempeño lector en las primeras etapas de aprendizaje formal (Ysla, 2015). En su estudio, la mayor atención ha recaído en las habilidades fonológicas, cuyo papel entre el primer y tercer año de primaria resulta ser clave (Parrila, Kirby y McQuarrie, 2004). La descripción de cómo los niños aprenden a manipular mentalmente las unidades del lenguaje escrito ha sido ampliamente abordado (Aguilar, Marchena, Navarro, Menacho y Alcalde, 2011; Casillas y Goikoetxea, 2007; Goikoetxea, 2005; Herrera y Defior, 2005), determinándose su carácter evolutivo y madurativo (Sellés y Martínez, 2014) y observándose una gradación en el dominio de los segmentos sonoros del habla y la complejidad de las tareas (Defior y Serrano, 2011). Todo ello es resultado de las experiencias con el lenguaje oral.

Otro predictor reconocido es el conocimiento alfabético. Junto a las habilidades fonológicas permite a los niños establecer una correspondencia sonido-letra. Esto facilita la decodificación del texto escrito. Para manejar el principio alfabético los niños deben comprender la correspondencia entre las letras impresas y los sonidos y para ello necesitan la representación fonéticamente estructurada del habla tanto como el conocimiento del sonido de la letra (Hulme, Bowyer-Crane, Carroll, Duff y Snowling, 2012). Se trata entonces de predictores relacionados. Incluso se ha afirmado que la conciencia fonológica es un predictor del conocimiento de las letras (Diuk y Ferroni, 2011). Ambos se presentan como facilitadores en la lectura durante los primeros años escolares (Schatschneider, Fletcher, Francis, Carlson y Foorman, 2004). Se ha encontrado que al iniciar el primer año escolar los niños que entraban conociendo algunas letras se desempeñaban mejor en la lectura en los primeros años de enseñanza primaria (Bravo, Villalón y Orellana, 2006), sobre todo en procesos como la decodificación. Sin embargo, no se ha identificado un rol directo en la comprensión de textos (Bowyer-Crane, Snowling, Duff, Fieldsend, Carroll, Miles y Hulme, 2008). Por esta razón, surge el interés por indagar y aportar evidencias respecto a otros precursores como el conocimiento metalingüístico. Se trata de otro predictor de la lectura, que en niños prelectores se pone de manifiesto con la toma de conciencia del entorno impreso y sus componentes. Consideramos que su estudio y promoción desde etapas tempranas aportaría al desarrollo de la competencia lectora, puesta en evidencia cuando el lector comprende, utiliza, reflexiona y se compromete con textos escritos, apoyado en una serie de destrezas que se van adquiriendo desde la escolaridad (OECD, 2010). Esta no solo se alcanza al decodificar y comprender lo que dice el texto, sino también con la puesta en marcha de procesos reflexivos. Se constituye por tanto en un indicador predictivo del éxito lector y si se detectase alguna dificultad en dicha variable indicaría también un riesgo para el aprendizaje lector (Romero y Castaño, 2016). Resulta esencial entonces plantear procedimientos de evaluación que permitan no solo reconocer su condición evolutiva sino también detectar de manera oportuna situaciones que dificulten la alfabetización formal.

1.1. El conocimiento metalingüístico

Este predictor está referido al conocimiento del niño acerca de la naturaleza del lenguaje escrito, sus formas y funciones (Pellicer y Baixauli, 2012). Es la toma de conciencia que se posee sobre las unidades que componen el lenguaje escrito -letra, palabra y frase- y el conocimiento sobre los usos y funcionalidad de la lectoescritura (Sellés, 2006, 2008; Sellés y Martínez, 2008). Su aparición no se da de modo espontáneo, surge cuando se enfrenta a determinadas tareas que le hacen reflexionar sobre el lenguaje (Jiménez, Rodrigo, Ortiz y Guzmán, 1999), por ejemplo cuando se da cuenta que los materiales impresos cumplen un papel informativo. Su valor reside en que los niños alcanzan a comprender que lo impreso tiene un significado, aunque no se encuentren habilitados para decodificar la información (Kassow, 2006). Los niños lo ponen en evidencia por ejemplo cuando dicen cuál es el propósito de la lectura, indican los procedimientos que siguen para realizar este tipo de tareas y dan muestras de autorregulación (Flórez-Romero, Torrado, Mondragón y Pérez, 2003). Tal es el caso de niños que observan a los adultos de su entorno familiar leyendo un diario y se dan cuenta que lo hacen para conocer, por ejemplo, el resultado del partido de fútbol del día anterior. Si se les pregunta, pueden llegar a explicar cómo llegaron a esa conclusión.

El conocimiento metalingüístico presenta características evolutivas comparables a las encontradas en el conocimiento fonológico, aunque habría que considerar que para la toma de conciencia en el ámbito metalingüístico es necesaria la intervención del adulto. En la etapa previa a la alfabetización formal es posible observar el manejo de nociones acerca de la direccionalidad del texto, la diferenciación entre las conductas de leer y escribir (alrededor de los 3 y 4 años), el reconocimiento de en qué consiste leer y escribir y de los propósitos o funciones del lenguaje escrito (5 años) hasta observar conocimientos sobre aspectos convencionales del lenguaje escrito (entre 5 y 6 años) (Ortiz y Jiménez, 2001). Parece ser que los niños adquieren primero la conciencia de lo escrito como actividad y luego descubren la noción de las palabras (Justice y Ezell, 2001). Esta habilidad que coloca al niño como conductor de sus propios procesos lectores podría ayudarle, por ejemplo a discriminar si se trata de una palabra o pseudopalabra y decidir si debe seguir la ruta fonológica o léxica en la tarea de leer.

El ser partícipe de actividades narrativas guiadas por el adulto podrían explicar mejor el desarrollo de este conocimiento. La exposición directa a una determinada cantidad de material de por sí no sería un predictor potente del aprendizaje de la lectura (Dickinson y Snow, 1987). El acceso y manipulación del material impreso (no solo cuentos, también materiales cotidianos como folletos publicitarios, periódicos, etc.) sí tendría consecuencias favorables. No solo es importante que el niño cuente en el hogar y la escuela con una buena cantidad de recursos bibliográficos, también es necesario promover actividades que hagan posible la toma de conciencia como por ejemplo compartir la lectura con un adulto. Esto serviría para la exploración del material por parte del niño y para que el adulto promueva habilidades tempranas de lectura (Neumann, Hood, Ford y Neumann, 2011). El resultado de esta interacción entre el ambiente, el papel del adulto y la maduración de los niños podría ser observado a partir de procedimientos que midan la toma de conciencia de lo impreso.

1.2. La evaluación del conocimiento metalingüístico

Los trabajos sobre la caracterización y evaluación de este predictor cuentan con mayores evidencian en poblaciones de habla inglesa que en lengua española. En este apartado desarrollaremos los procedimientos que se han seguido en la medición de este predictor en preescolares. En estudios realizados con niños, cuyo idioma es el inglés, se han empleado tareas como la presentación del nombre de objetos que son propios de su comunidad para medir conciencia del material impreso (Vera, 2011). Justice y Ezell (2001) evaluaron en niños, de edad promedio 4 años, los conocimientos sobre el material impreso. Con la ayuda de cuentos, los niños debían realizar tareas tales como señalar las letras mayúsculas, o indicar el principio del texto, lo que llevó a observar que se les hacía más sencillo identificar palabras según su extensión (larga o corta), pero se les complicaba encontrar la última palabra de un texto e identificar la letra capital. En cuanto al concepto de lo impreso estos autores observaron que los niños identificaban el título del libro (quizá porque iba acompañado de una imagen y estaba en la portada), la portada, la orientación de la lectura (de izquierda a derecha), la primera línea del texto e intuyeron qué decía el título. Justice, Bowles y Skibbe (2006) incluyeron en la evaluación de la conciencia de lo impreso en niños en riesgo entre los 3 y 5 años, tareas tales como distinguir entre letras y figuras, identificar el título de un libro y el conocimiento del alfabeto.

Algunas baterías en inglés dirigidas a evaluar las habilidades prelectoras también han incluido tareas que miden este conocimiento. El *test of Preschool Emergent Literacy – TOPEL* de Lonigan, Wagner, Torgesen y Rashotte (2007) va dirigido a niños de edades comprendidas entre los 3 y los 5 años y 11 meses es una ellas. Considera en la evaluación de la conciencia de lo impreso: el conocimiento del alfabeto así como el reconocimiento de letras y palabras impresas. Contiene tareas como señalar la figura que tiene una palabra en ella, discriminar palabras e identificar letras, el nombre y su sonido, apoyándose en estímulos visuales.

Otra herramienta de medición de las habilidades prelectoras que incluyó en su evaluación el conocimiento de lo impreso es el *Get Ready To Read* (Whitehurst y Lonigan, 2001). Se emplea en niños preescolares entre 3 y 5 años. La versión más reciente contiene 25 ítems, los que evalúan la conciencia de lo impreso y la conciencia fonológica. En todos los casos se apoyan en material visual y, con la orientación del examinador, permite valorar si el niño logra distinguir las letras y palabras de entre otros estímulos (como números o símbolos) así como ciertas características de un libro (la extensión de un cuento o la contraportada). Ha sido empleada junto al *TOPEL* en estudios para determinar su capacidad valorativa en los predictores de la lectura (Lonigan, Allan y Lerner, 2011; Wilson y Lonigan, 2009).

En cuanto a poblaciones de habla española, el instrumento que ha incluido entre sus tareas la evaluación de la toma de conciencia de lo impreso (y que es empleado en el presente estudio) es la Batería de Inicio a la Lectura-BIL (Sellés, Martínez, Vidal-Abarca y Gilabert, 2008). Su objetivo es evaluar las habilidades cognitivas y lingüísticas relacionadas con el éxito en el aprendizaje inicial de la lectura, y estimar en qué grado se encuentran desarrollados 5 factores en la etapa previa a la alfabetización formal. Uno de esos factores es el conocimiento metalingüístico sobre la lectura con tres pruebas: reconocer palabras, reconocer frases e identificar las funciones de la lectura. Su aplicación se realiza con apoyo de material visual. En un estudio con niños peruanos que

asisten a centros públicos (Ysla, 2015) se encontró que ciertas características demográficas influyen en su desarrollo. Estos factores fueron la edad (a favor de los mayores), el sexo (con mejores puntuaciones para las niñas) y el número de libros en el hogar (en el que niños que contaban entre 25 y 100 libros presentaban un mejor desempeño).

Estos avances en evaluación incentivan la búsqueda de mayores evidencias en poblaciones de habla española. La descripción de estas tareas sin embargo debe considerar el contexto educativo en el que se presenta por lo que se desarrolla brevemente el enfoque de enseñanza de la escuela peruana.

1.3. El currículo escolar peruano

La educación básica regular del Perú se organiza en tres niveles educativos: educación inicial (infantil), primaria y secundaria. Esta se rige por un currículo orientado bajo el enfoque por competencias, y desde el que se articulan los aprendizajes a desarrollar durante toda la etapa escolar. En lo que respecta a la competencia lectora, el propósito es que al finalizar la educación secundaria los estudiantes obtengan, infieran e interpreten información y reflexionen sobre la forma, el contenido y contexto de diversos tipos de textos escritos (Ministerio de Educación del Perú, 2016). La alfabetización se inicia formalmente cuando los niños ingresan al primer grado de primaria, por lo que en la etapa previa (educación infantil) se espera que los niños "lean" por sí mismos (sin haber adquirido el sistema de escritura alfabética) textos en variadas situaciones comunicativas y elaboren diversas hipótesis sobre lo que estos dicen, llegando a una comprensión crítica y relacionando sus conocimientos previos con los elementos que reconocen en los textos: imágenes, indicios, palabras o letras, entre otros (Ministerio de Educación de Perú, 2013). Es por ello que en las aulas de este nivel educativo se involucra a los niños en actividades de lectura y escritura de manera más natural, dejándoles la posibilidad que descubran por sí mismos, con orientación de su docente, las características del material escrito que manipulan.

Las acciones pedagógicas de este aprendizaje se conducen desde el enfoque comunicativo textual. Comunicativo porque promueve el desarrollo de conocimientos y destrezas necesarias para utilizar eficazmente el lenguaje en situaciones concretas de la vida, a partir de una didáctica basada en la reflexión y el análisis acerca de oraciones, textos, diálogos y otras unidades lingüísticas enunciadas en situaciones comunicativas (Ministerio de Educación de Perú, 2015). Esto implica desarrollar una serie de habilidades que le permitan acceder a la información e interactuar con ella. Por el lado de lo textual, se asume que la producción escrita es la unidad lingüística fundamental de comunicación y comprende una interacción entre el que escribe y quien lee, y adquiere sentido cuando se pone en contacto la información que contiene con el conocimiento que maneja el lector. Estos niveles podrían ejemplificarse desde la enseñanza en educación infantil con el nombre del niño: desde lo comunicativo, lo reconoce impreso, en sus pertenencias personales, uniforme o en un cartel de asistencia; visto desde lo textual, sabe que esa combinación de letras representa su nombre y se emplea justamente para distinguirlo del resto de personas. Los planteamientos de este enfoque parecen ir en concordancia con el desarrollo del conocimiento metalingüístico, dado que apuesta por la exposición a experiencias que posicionen al niño en un mundo escrito. Para concretar esta intencionalidad pedagógica se requieren mayores elementos de análisis respecto a los procesos reflexivos que es capaz de promover el niño en la etapa infantil. Por su

parte, el análisis de las unidades sonoras del habla se aborda a partir de la oralidad, sin necesariamente establecer un vínculo con el sistema de escritura. Este proceso más bien se ha previsto ser trabajado al iniciar la educación primaria.

Teniendo en cuenta entonces la orientación pedagógica del currículo peruano, desde el presente estudio se propone evaluar el conocimiento metalingüístico en estudiantes del último ciclo de educación infantil de un centro público peruano. Nos interesa conocer en qué situación se encuentra este predictor, valorar el trabajo de aula, el cual se orienta bajo un currículo que enfatiza más el desarrollo de la oralidad en la educación preescolar. Esta información nos permitiría plantear un estudio a mayor profundidad, ampliando el rango de edad (en vías de una investigación longitudinal). Asimismo, al encontrarse próximos a la enseñanza formal de la lectura, resulta necesario no solo caracterizar el conocimiento metalingüístico, sino también detectar alguna situación que ponga en riesgo dicho aprendizaje. Para este trabajo se ha empleado el BIL, dirigida a poblaciones de habla española, con tres tareas que miden este predictor: reconocer palabras, reconocer frases e identificar las funciones de la lectura.

2. Método

El estudio consiste en describir el desarrollo del conocimiento metalingüístico en niños que asisten a aulas del último ciclo de la educación infantil peruana, antes de la enseñanza formal de la lectura. Se trata de un trabajo netamente descriptivo, sin ningún tipo de intervención, y con una única medición.

La variable de estudio es el conocimiento metalingüístico, medido a partir de tres tareas: reconocer palabras, reconocer frases e identificar las funciones de la lectura. Al tratarse de un predictor que favorece el aprendizaje inicial de la lectura, consideramos importante reconocer cómo se evidencian estas tareas en esta etapa del desarrollo infantil.

2.1. Participantes

Formaron parte del estudio 90 niños del último ciclo de educación infantil de un centro público peruano. Previamente se consultó a otras instituciones, siendo esta la escuela que brindaba las facilidades y condiciones de espacio y tiempo para realizar la evaluación. Antes de iniciar el trabajo se confirmó que la institución no realizara una intervención específica en las habilidades de inicio a la lectura.

Esta escuela, con más de 30 años de funcionamiento, se encuentra ubicada en el distrito de San Luis (ciudad de Lima), cuya incidencia de pobreza es baja, de acuerdo a los últimos reportes del Instituto Nacional de Estadística e Informática – INEI (2015). El equipo de docentes cuenta con más de 10 años de experiencia en el sector y nivel educativo. Al igual que otras instituciones de educación infantil peruana, el equipo de profesoras ha recibido como orientación promover actitudes en los niños hacia la lectura a partir de actividades cotidianas. No consideran entre sus estrategias trabajar el análisis de los elementos sonoros del habla.

Fueron evaluados 58 niños (64,4%) y 32 niñas (35,6%) quienes, de acuerdo al informe brindado por sus docentes, no se encontraban en riesgo de dificultades de aprendizaje. Si bien se encontraban matriculados en el aula de 5 años, fueron distribuidos de acuerdo a los grupos de edad propuestos en la Batería de Inicio a la Lectura-BIL (por meses). Esto nos ayudaría a especificar si con el paso de los meses se evidencian diferencias en los

desempeños. De acuerdo a la fecha en la que se aplicó el instrumento, a mediados de curso escolar, los niños fueron organizados en tres grupos: el de 5 años hasta 5 años y 5 meses contaba con 8 niños (8,9%), los que tenían entre 5 años y 6 meses y 5 años y 11 meses fueron 46 (51,1%) y el grupo de niños de 6 años a más estuvo conformado por 36 estudiantes (40 %) (Ver tabla 1).

Tabla 1. Distribución edad y sexo

EDAD	NIÑA		NIÑO		TOTAL	
	N	%	N	%	N	%
Entre 5 años y 5 años y 5 meses	5	5,6	3	3,3	8	8,9
Entre 5 años y 6 meses y 5 años y 11 meses	15	16,7	31	34,4	46	51,1
6 años a más	12	13,3	24	26,7	36	40
Total	*32*	*35,6*	*58*	*64.,4*	*90*	*100*

Fuente: Elaboración propia.

Muchos de los niños matriculados ingresaron posiblemente con más 5 años y 6 meses. Al aplicar el instrumento, en el mes de julio, un gran número de ellos ya había cumplido los 6 años y es por ello que se constituye en el segundo grupo con mayor número de participantes (36).

La escuela que formó parte de este estudio atiende en dos turnos: mañana (47 niños) y tarde (43 niños). Tanto en la mañana como en la tarde se cuenta con dos aulas (ver tabla 2).

Tabla 2. Distribución turno y aula

TURNO	NIÑOS EVALUADOS		AULA	NIÑOS EVALUADOS	
	N	%		N	%
Turno mañana	47	52,2	1	24	26,7
			2	23	25,6
Turno tarde	43	47,8	3	22	24,4
			4	21	23,3

Fuente: Elaboración propia.

2.2. Instrumento

Se empleó la Batería de Inicio a la Lectura-BIL para niños de 3 a 6 años (Sellés et al. 2008). El instrumento en su totalidad cuenta con 15 pruebas agrupadas en 5 factores y 143 ítems que miden además del conocimiento metalingüístico, el conocimiento fonológico; el conocimiento alfabético; las habilidades lingüísticas y los procesos cognitivos.

La prueba del conocimiento metalingüístico está compuesta por tres tareas. En la primera, *reconocer palabras*, se presentaba a los niños un listado de palabras y no palabras (10 estímulos impresos). Ellos debían indicar cuáles lo eran y cuáles no. En la segunda tarea, *reconocer frases*, similar a la tarea anterior, se les pedía indicar en un listado de cinco estímulos si eran frases o no. Por último, en la tarea *funciones de la lectura*, se les narraba una secuencia representada visualmente en cuatro imágenes en las que el niño debía indicar para qué le había servido leer al personaje. Fueron cinco las escenas trabajadas en esta evaluación. Todas las tareas propuestas se apoyaron en material impreso. Las respuestas proporcionadas por los niños quedaban registradas en un cuadernillo. Cada acierto correspondía un punto. Los errores no se contabilizaban. Para

el presente trabajo se trabaja con las puntuaciones directas obtenidas en cada una de las tareas.

Este instrumento ya ha sido empleado en otros estudios con población peruana (Ávila e Ysla, 2014; Ysla, 2015). Asimismo, la información que se trabaja con los niños no ha requerido de un proceso de contextualización puesto que se refiere básicamente a características del material impreso, donde no se hace referencia a un vocabulario ajeno al empleado en la ciudad de Lima.

2.3. Procedimiento

El trabajo con la escuela seleccionada exigió una serie de coordinaciones previas con la directora y profesoras de aula, para contar con su autorización y coordinar los horarios y tiempos que tomaría su aplicación. La aplicación se realizó a mediados de curso escolar, en el mes de julio, cuatro meses después de haber iniciado las clases.

El instrumento se aplicó individualmente. Para ello, la escuela dispuso de un espacio privado para las evaluaciones. La aplicación se realizó en el horario de clases. Uno a uno salía cada niño para realizar el trabajo con la examinadora. Esta parte de la batería tomó entre 8 a 10 minutos por niño. Las profesoras proporcionaron los datos como fecha de nacimiento, información que junto a la fecha de evaluación, permitiría determinar la edad exacta.

Con los datos obtenidos, se realizaron los estadísticos descriptivos de toda la muestra y por grupos de edad. Para establecer si el conocimiento metalingüístico presenta diferencias entre los grupos de edad, se utilizó en un primer momento el análisis de varianza (ANOVA). Sin embargo, al ser el tamaño de cada grupo diferente, se empleó la prueba Kruskal-Wallis (prueba no paramétrica) para confirmar dicha varianza. Todo el procesamiento de la información fue realizado con el paquete estadístico SPSS (versión 22).

3. Resultados

Se presentan los estadísticos descriptivos (media y desviación estándar) de las puntuaciones directas obtenidas en las tareas evaluadas. Con la sumatoria de las tres pruebas se obtuvo la puntuación total del conocimiento metalingüístico. También se muestran los datos alcanzados con la prueba Kruskal-Wallis, siendo la variable independiente la edad y las dependientes las tareas del conocimiento metalingüístico.

3.1. El conocimiento metalingüístico

Al revisar las puntuaciones obtenidas por el grupo de participantes (90 niños) se obtuvo una media de 10,61 (DT=2,91). Las diferencias entre edades fueron significativas (x^2 = 6,877; p= 0,032; prueba Kruskal-Wallis).

El grado de dificultad de las tareas se hace menor conforme pasan de un grupo de edad al otro. Los niños entre los 5 años y los 5 años y 5 meses obtienen una media de 8,63 (DT=2,75), los que tienen entre 5 años y 6 meses y 5 años y 11 meses 10,30 (DT=2,94) y los de 6 años a más alcanzan una media de 11,43 (DT=2,69). Encontramos entonces que la capacidad de los niños para tomar conciencia de las características del material impreso (reconocimiento de palabras y de frases y la identificación de las funciones de la lectura) es posible de ser observada en este grupo de edad.

3.2. Reconocimiento de palabras

La tarea contemplaba distinguir cuáles eran palabras y cuáles no en una lista de 10 que contenía palabras y otros estímulos (como números o símbolos). Para ello recibían la indicación: *"Me vas a decir si lo que ves es una palabra o no. ¿Es una palabra? ¿Sí o no?" A* cada respuesta correcta se le asignaba un punto. Antes de realizar el ejercicio se les presentaba un ejemplo, para asegurar que habían comprendido la tarea. Al revisar los resultados se observa que la muestra total obtuvo una media de 7,72 (DT=2,24). Es decir, de 10 estímulos los niños lograban decir casi 8 respuestas correctas.

Entre los estímulos de mayor facilidad se encontraban aquellos que representaban el nombre de personas (como Rosa y Laura) o los que contenían algún número o símbolo (los niños inmediatamente los descartaban y algunos decían que no eran palabras sino números). Los estímulos que generaban confusión eran las palabras presentadas en mayúsculas.

Al realizar el análisis de varianza se identificó una diferencia significativa entre los tres grupos de edad (x^2 = 6,462; p= 0,040; prueba Kruskal-Wallis). Nuevamente, los niños mayores, el grupo de 6 años a más, demostraron mejores desempeños (M=8,19; DT=2,10) a diferencia de los menores. Los niños entre 5 años y 5 años y 5 meses alcanzaron una media de 5.75 (DT=2,60) mientras que el grupo entre 5 años y 6 meses y 5 años y 11 meses obtuvo una media de 7.70 (DT=2,10). A pesar que los niños de 6 años serían el último grupo a quien va dirigida la escala BIL, encontramos que no hay efecto techo en este grupo.

De acuerdo a estos resultados, la toma de conciencia de las palabras y su distinción con las no palabras sugiere una evolución asociada a la edad de los niños. En este caso las diferencias son significativas entre los tres grupos. Los niños mayores cometen menos errores en tareas que ponen a prueba el reconocimiento de las palabras.

3.3. Reconocimiento de frases

En esta tarea los niños debían decidir si el elemento que se les presentaba era una frase. Fueron 5 los estímulos visuales y bajo la indicación: *"¿Es esto una frase?"* Los niños debían observar y decidir sí o no. Al igual que la tarea anterior, primero debían resolver el ejemplo presentado por la examinadora. A cada respuesta correcta se le asignaba un punto. El máximo puntaje a alcanzar es 5. El grupo total (90 niños) alcanzó una media de 3,67 (DT=1,64) y, a diferencia del *reconocimiento de palabras*, no fue posible encontrar diferencias significativas entre los grupos (x^2 =,382; p= 0,826; prueba Kruskal-Wallis) aun cuando los niños mayores presentaron mejores desempeños.

Uno de los estímulos de mayor facilidad contenía no solo palabras y números sino también símbolos o dibujos. Con este rasgo los niños lo descartaban inmediatamente. Los que generaban mayor confusión contenían signos de interrogación y/o exclamación. Parece ser que al manipular materiales escritos (como cuentos) en los que aparecen señalizados diálogos con estos signos, los niños lo asocian directamente con frases. Sin embargo, uno de los estímulos contenía palabras rodeadas de signos de exclamación e interrogación, y aunque no representaban una idea, los niños lo daban como válido (podemos pensar que es por el hecho de contener estos elementos).

Al revisar las puntuaciones, encontramos similares desempeños. Así, el grupo de 6 años a más obtiene una media de 3,78 (DT=1,59), el de los niños entre 5 años y 6 meses y 5

años y 11 meses alcanza una media de 3,61 (DT=1,71) y los menores (entre 5 años y 5 años y 5 meses) alcanzan una media de 3,50 (DT=1,70).

3.4. Funciones de la lectura

En esta última tarea los niños debían demostrar un conocimiento aproximado de para qué sirve leer. Para tal fin, la examinadora les narró 5 historias, cada una apoyada con cuatro viñetas, en las que los personajes leían con un propósito diferente. Conforme se iba narrando la historia se iba señalando cada viñeta. Al final, se les preguntaba*: "¿Para qué le ha servido leer?"* Las respuestas de los niños eran anotadas en un cuadernillo y al final se cotejaba con los propósitos definidos en el manual: (1) para recordar la información en otro momento, (2) por entretenimiento, (3) y (5) para informarse y (4) para adquirir conocimiento. De las 5 narraciones, el grupo alcanzó una media de 3,08 (DT=1,22). Aquí también ha sido posible encontrar diferencias significativas entre los tres grupos de edad (x^2 =11,296; p= 0,004; prueba Kruskal-Wallis) siempre a favor de los niños mayores.

Esta es sin duda una de las tareas que revela directamente la noción de los niños respecto al papel de la lectura en situaciones cotidianas. Uno de las historias de las que fácilmente rescataban la función de la lectura (por entretenimiento) trataba de cómo un grupo de niños participaba en la narración de un cuento. Las de mayor dificultad se referían al propósito de leer un periódico y la revisión de un folleto publicitario (para informarse).

Fueron los niños de 6 años los que obtuvieron mejores puntuaciones (M=3,56; DT=1,11) seguidos del grupo entre 5 años y 6 meses y 5 años y 11 meses (M=2,85; DT=1,21). Fue el grupo de menor edad el que obtuvo las puntuaciones más bajas (M=2,25; DT=1,03).

Los resultados evidencian que las tareas propuestas en el BIL parecen sensibles a la evolución y maduración de los niños que asisten al último año de educación infantil, bajo un currículo que apunta a un enfoque contextual.

4. Discusión y conclusiones

El proceso seguido nos permite, en primer lugar, confirmar que se trata de un predictor observable en etapas iniciales del desarrollo lector, cuando el niño empieza a entrar en contacto con el material impreso y previo a la escolarización obligatoria. Esta información contribuye a un mayor reconocimiento de esta habilidad antes de la alfabetización formal, más aun considerando que se trata de una habilidad prelectora que requiere la facilitación del adulto, en este caso, de la docente de aula. La evidencia encontrada también puede ser empleada para orientar la implementación de acciones pedagógicas específicas en el aula o intervenir en situaciones de riesgo.

Los niños de este estudio demuestran que son conscientes que el material impreso está compuesto de palabras y frases y que además cumple diversas funciones, y que un mayor dominio está asociado a la edad. La información obtenida a partir de las puntuaciones alcanzadas tanto en el conocimiento metalingüístico como en las tareas *reconocer palabras* y *funciones de la lectura* así lo confirma. Este desarrollo se da dentro de un contexto específico de aprendizaje. Por ello consideramos importante describir la orientación del currículo peruano, que si bien en esta etapa enfatiza el desarrollo del componente oral

del lenguaje infantil, procura introducir a los niños en la manipulación y exploración del mundo escrito de modo más inductivo. Aunque no se ha medido directamente el papel de las docentes, podemos sospechar que el desarrollo de este predictor es resultado de la combinación de experiencias y la maduración y no estaría limitado a la sola exposición del material impreso (Dickinson y Snow, 1987). Consideramos que podría ser la interacción del niño con la docente, al manipular y reconocer las características del material impreso (como cuentos, diarios, carteles, entre otros), el punto de partida para que empiece a reflexionar sobre el mundo escrito. Sería interesante realizar estudios sobre qué estrategias emplean las docentes para promover esta interacción.

Los resultados respecto al *reconocimiento de las palabras* ponen en evidencia que el niño tiene conciencia de la palabra (Niessen et al., 2011). A esta edad es capaz de reconocer que el lenguaje escrito está compuesto de unidades y de comprender que la palabra expresa un significado (Kassow, 2006). Esto podría estar asociado a otro predictor, como el conocimiento de letras. Esto lo observamos cuando los niños justificaban que no se trataba de una palabra cuando lo que se le presentaban eran números ("estos son números, no puede ser una palabra") o cuando sin saber leer se daban cuenta que era el nombre de una persona ("Rosa"). Este tipo de tareas obliga al niño a regularse y pensar antes de dar una respuesta. Muchos niños hacían una pequeña pausa antes de decir si se trataba o no de una palabra. Esto reflejaría también una actitud reflexiva frente a este tipo de tareas. Hay un factor clave de autorregulación y conducción que se pone en marcha (Flórez-Romero et al., 2003). Para reconocer y distinguir las palabras entre estímulos que además contienen números y símbolos, el niño observa y discrimina si se trata de un conjunto de letras o si son cifras o signos. Ese es quizá uno de los principios que se pone en evidencia para realizar la tarea. Esto nos llevaría a inferir que, la autorregulación se evidencia antes del aprendizaje de la lectura.

Al no haberse encontrado diferencias significativas en el *reconocimiento de frases*, podría inferirse que la toma de conciencia de este aspecto requiere una mayor intervención por parte del adulto, tanto del padre de familia como del responsable del aula. Los tres grupos demuestran haber tenido similar dificultad para distinguir cuando se enfrentaban a una frase o a un conjunto de símbolos. Una tarea previa podría estar relacionada a las evidencias proporcionadas por Justice y Ezell (2001) quienes encontraron que a la edad promedio de 4 años los niños no solo podían identificar en la portada de un cuento el título, sino también intuir de qué se trataba el título del mismo

Respecto al conocimiento de las *funciones de la lectura*, en cada una de las narraciones el niño debía identificar para qué le sirvió leer al personaje de la historia. Se trataría de un nivel de toma de conciencia en la que el niño asume el papel que cumple la lectura en situaciones cotidianas. Al respecto, se ha señalado que los niños a la edad de cinco años pueden reconocer en qué consiste leer y los propósitos y funciones del lenguaje escrito (Ortiz y Jiménez, 2001). Esto representa un nivel de comprensión respecto al rol del lector frente al texto. En este caso se trataba de ponerse en el lugar del personaje para reconocer la utilidad del acto de leer. Los casos no solo se diferenciaban por la situación a la que se enfrentaba el personaje sino también porque en cada historia se hacía alusión a un tipo de material (folleto, periódico, libro, nota de apuntes, cuento). Otra vez la experiencia y el acceso a material diverso podrían haber incidido en el logro de esta tarea. Si un niño a esta edad no ha manipulado e interactuado con estos materiales y reconocido su uso, difícilmente sabría indicar para qué le ha servido leer al personaje. Vemos que ha sido más reconocida la función de diversión unida a la lectura de cuentos

(actividad que les resulta más familiar a los niños de esta edad). Sin embargo, no ocurre lo mismo con aquellas tareas como leer un periódico o revisar un folleto publicitario. Les es más difícil identificar el propósito puesto que no se encuentran involucrados de manera directa en este tipo de actividades. Sus referencias directas serían los adultos que tienen prácticas lectoras en el hogar.

La evaluación de este predictor podría también incorporar otros componentes, como por ejemplo señalar en un texto las letras mayúsculas, el principio del mismo, la letra capital (con la que empieza) o la última palabra (Justice y Ezell, 2001). Así, se contarían con mayores elementos para marcar una secuencia de cómo se complejiza el conocimiento metalingüístico.

El haber elegido para esta evaluación a un único grupo de edad (5 años) ha implicado una limitación. De acuerdo a la distribución de los grupos propuestos por la BIL, no se parte de grupos homogéneos en cantidad. Esto en razón de la complejidad del sistema educativo peruano y el propio contexto. Por ello es relevante continuar con estos procesos de indagación que abarquen las otras edades de la educación infantil (3, 4 y 5 años). El paso de un grupo a otro supone un avance en los procesos reflexivos que ponen en marcha para identificar componentes del lenguaje escrito como son las palabras y el propósito de la lectura. Si el estudio se ampliara a una muestra con estas edades podría confirmarse su condición evolutiva. Esta información resultaría de gran ayuda para los responsables del currículo escolar y de las acciones pedagógicas, quienes contarían con mayores elementos a fin de proporcionar experiencias que apuesten por la reflexión y toma de conciencia del mundo escrito, de acuerdo a las posibilidades de cada niño. Nuevos trabajos podrían valorar qué tanto se concreta en el trabajo de aula el desarrollo de la toma de conciencia de lo escrito, más en un contexto peruano que ha enfatizado el desarrollo de la oralidad en la etapa preescolar.

Por último, para el presente estudio se ha considerado como variable independiente la edad de los participantes evaluados. Futuros trabajos podrían incluir otros factores que favorecen o inciden en el desarrollo del conocimiento metalingüístico. De este modo podría llegarse a una mayor profundización sobre este predictor. El cruce de los resultados podría hacerse también con variables sociodemográficas como el sexo, el lugar de vivienda o el nivel educativo y laboral de los padres de familia. También podrían considerarse las interacciones niño-adulto con el material impreso en el hogar. Esto permitiría confirmar que si bien hay una disposición madurativa por parte de los niños, existirían otros factores, contextuales, que favorecerían el desempeño en estas tareas.

Referencias

Aguilar, M., Marchena, E., Navarro, J., Menacho, I. y Alcalde, C. (2011). Niveles de dificultad de la conciencia fonológica y aprendizaje lector. *Revista de Logopedia, Foniatría y Audiología, 31*(2), 96-105. doi:10.1016/s0214-4603(11)70177-2

Ávila, V. e Ysla, L. (2014, septiembre). El papel de la familia, escuela y contexto en el vocabulario y la gramática infantil: El caso de niños peruanos y españoles. Póster presentado en el *XXIX Congreso Internacional AELFA*, Universidad de Murcia.

Bowyer-Crane, C., Snowling, M. J., Duff, F. J., Fieldsend, E., Carroll, J. M., Miles, J., Götz, K. y Hulme, C. (2008). Improving early language and literacy skills: Differential effects of an

oral language versus a phonology with reading intervention. *Journal of Child Psychology and Psychiatry*, *49*(4), 422–432. doi:10.1111/j.1469-7610.2007.01849.x

Bravo, L., Villalón, M. y Orellana, E. (2006). Predictibilidad del rendimiento en la lectura: Una investigación de seguimiento de primer a tercer año. *Revista Latinoamericana de Psicología*, *38*(1), 9-20.

Casillas, A. y Goikoetxea, E. (2007). Sílaba, principio-rima y fonema como predictores de la lectura y la escritura tempranas. *Infancia y Aprendizaje*: *Journal for the Study of Education and Development*, *30*(2), 245-259. doi:10.1174/021037007780705184

Defior, S. y Serrano, F. (2011). La conciencia fonémica, aliada de la adquisición del lenguaje escrito. *Revista de Logopedia*, *Foniatría y Audiología*, *31*(1), 2-13. doi:10.1016/s0214-4603(11)70165-6

Dickinson, D. K. y Snow, C. E. (1987). Interrelationships among prereading and oral language skills in kindergartners from two social classes. *Early Childhood Research Quarterly*, *2*(1), 1-25. doi:10.1016/0885-2006(87)90010-x

Diuk, B. y Ferroni, M. (2011). Predictors of letter knowledge in children growing in poverty. *Psicologia: Reflexão e Crítica*, *24*(3), 570-576. doi:10.1590/s0102-79722011000300018

Flórez-Romero, R., Torrado, M. C., Mondragón, S. P. y Pérez, C. (2003). Explorando la metacognición: Evidencia en actividades de lectura y escritura en niños y niñas de 5 a 10 años de edad. *Revista Colombiana de Psicología*, *12*, 85-98.

Goikoetxea, E. (2005). Levels of phonological awareness in preliterate and literate Spanish-speaking children. *Reading and Writing*, *18*(1), 51–79. doi:10.1007/s11145-004-1955-7

Herrera, L. y Defior, S. (2005). Una aproximación al procesamiento fonológico de los niños prelectores. Conciencia fonológica a corto plazo y denominación. *Psykhe*, *14*(2), 81-95. doi:10.4067/s0718-22282005000200007

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J. y Snowling, M. J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read: Combining intervention studies with mediation analyses. *Psychological Science*, *23*(6), 572-576. doi:10.1177/0956797611435921

Instituto Nacional de Estadística e Informática (INEI). (2015). *Mapa de pobreza provincial y distrital 2013*. Lima: Autor.

Invernizzi, M., Sullivan, A., Meier, J. y Swank, L. (2004). *Phonological Awareness Literacy Screening*: *Preschool (PALS-PreK)*. Charlottesville, VA: University of Virginia.

Jiménez, J.E., Rodrigo, M., Ortiz, M. y Guzmán, R. (1999). Procedimientos de evaluación e intervención en el aprendizaje de la lectura y sus dificultades desde una perspectiva cognitiva, *Infancia y Aprendizaje: Journal for the Study of Education and Development*, *22*(88), 107-122. doi:10.1174/021037099760246644

Justice, L. M., Bowles, R. P. y Skibbe, L. E. (2006). Measuring preschool attainment of print-concept knowledge: A study of typical and at-risk 3- to 5-year-old children using item response theory. *Language*, *Speech*, *and Hearing Services in Schools*, *37*(3), 224–235. doi:10.1044/0161-1461(2006/024)

Justice, L. M. y Ezell, H. K. (2001). Word and print awareness in 4 year-old children. *Child Languaje Teaching and Therapy*, *17*(3), 207-225. doi:10.1191/026565901680666527

Kassow, D. Z. (2006). Environmental print awareness in young children. *Talaris Research Institute*, *1*(3), 1-8.

Lonigan, C. J., Allan, N. P. y Lerner, M. D. (2011). Assessment of preschool early literacy skills: Linking children's educational needs with empirically supported instructional activities. *Psychology in the Schools*, *48*(5), 488–501. doi:10.1002/pits.20569

Lonigan, C. J., Wagner, R. K., Torgesen, J. K. y Rashotte, C. (2007). *The Test of Preschool Early Literacy*. Austin, TX: Pro-Ed.

Marder, S. E. (2008). *Impacto de un programa de alfabetización temprana en niños de sectores urbano marginals* (Tesis de posgrado). Universidad Nacional de La Plata, Argentina.

Ministerio de Educación de Perú. (2016). *Currículo Nacional de la Educación Básica*. Lima: Autor. Recuperado de http://www.minedu.gob.pe/curriculo/pdf/curriculo-nacional-2016-2.pdf

Ministerio de Educación de Perú. (2013). *Rutas del aprendizaje. ¿Qué y cómo aprenden nuestros niños y niñas? Fascículo 1. Desarrollo de la Comunicación. 3, 4 y 5 años de Educación Inicial*. Lima: Autor.

Ministerio de Educación de Perú. (2015). *Rutas del aprendizaje. Versión 2015. ¿Qué y cómo aprenden nuestros niños y niñas? Fascículo 1. Desarrollo de la Comunicación. 3, 4 y 5 años de Educación Inicial*. Lima: Autor.

Neumann, M. (2013). Using environmental print to foster emergent literacy in childrenfrom a low-SES community. *Early Childhood Research Quarterly*, *29*(3), 310-318. doi:10.1016/j.ecresq.2014.03.005

Neumann, M. M., Hood, M., Ford, R. M. y Neumann, D. L. (2011). The role of environmental print in emergent literacy. *Journal of Early Childhood Literacy*, *12*(3) 231–258. doi:10.1177/1468798411417080

Niessen, N. L., Strattman, K. y Scudder, R. (2011). The influence of three emergent literacy skills on the invented spellings of 4-year-olds. *Communication Disorders Quarterly*, *32*(2) 93–102. doi:10.1177/1525740110363624

OECD. (2010). *PISA 2009 assessment framework key competencies in reading, mathematics and science*. París: OECD Publishing. doi:10.1787/9789264062658-en

Ortiz, M. y Jiménez, J. E. (2001). Concepciones tempranas acerca del lenguaje escrito en prelectores. *Infancia y Aprendizaje: Journal for the Study of Education and Development*, *24*(2), 215-231. doi:10.1174/021037001316920744

Parrila, R., Kirby, J. y McQuarrie, L. (2004). Articulation rate, naming speed, verbal short-term memory and phonological awareness: Longitudinal predictors of early reading development? *Scientific Studies of Reading*, *8*(1), 3-26. doi:10.1207/s1532799xssr0801_2

Pellicer, A. y Baixauli, I. (2012). Intervención preventiva en las dificultades de la lectura y la escritura. *Boletín de AELFA*, *12*(2), 67-75. doi:10.1016/s1137-8174(12)70064-x

Romero, A. y Castaño, C. (2016). Prevenir las dificultades lectoras: Diseño y evaluación de un software educativo. *Píxel-Bit: Revista de Medios y Educación*, *49*, 207-223. doi:10.12795/pixelbit.2016.i49.14

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D. y Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, *96*(2), 265–282. doi:10.1037/0022-0663.96.2.265

Sellés, P. y Martínez, T. (2014). Secuencia evolutiva del conocimiento fonológico en niños prelectores. *Revista de Logopedia, Foniatría y Audiología*, *34*(3), 118-128. doi:10.1016/j.rlfa.2013.09.001

Sellés, P., Martínez, T., Vidal-Abarca, E. y Gilabert, R. (2008). *Manual Batería de Inicio a la Lectura para niños de 3 a 6 años (BIL 3-6)*. Madrid: ICCE.

Sellés, P. y Martínez, T. (2008). Evaluación de los predictores y facilitadores de la lectura: Análisis y comparación de pruebas en español y en inglés. *Bordón, 60*(3), 113-129.

Sellés, P. (2008). *Elaboración de una prueba de habilidades relacionadas con el desarrollo inicial de la lectura (BIL 3-6)* (Tesis doctoral). Universidad de Valencia, Valencia.

Sellés, P. (2006). Estado actual de la evaluación de los predictores y de las habilidades relacionadas con el desarrollo inicial de la lectura. *Revista Aula Abierta, 88*, 53-72.

Vera, D. (2011). Using popular culture print to increase emergent literacy skills in one high-poverty urban school district. *Journal of Early Childhood Literacy, 11*(3) 307–330. doi:10.1177/1468798411409297

Whitehurst, G. J. y Lonigan, C. J. (2001). *Get ready to read! Screening tool.* Nueva York, NY: National Center for Learning Disabilities.

Wilson, S. B. y Lonigan, C. J. (2009). An evaluation of two emergent literacy screening tools for preschool children. *Annals of Dyslexia, 59*(2), 115–131. doi:10.1007/s11881-009-0026-9

Ysla, L. (2015). *La intervención en las habilidades de inicio a la lectura en la educación infantil y su relación con los procesos lectores en niños de primer grado de primaria* (Tesis doctoral). Universidad de Valencia, Valencia.

Breve CV de los autores

Liz Cristina Ysla Almonacid

Docente de educación superior con estudios posgrado en Psicología del Desarrollo y Trastornos de la Comunicación Humana. Tiene a cargo cursos de Investigación-Acción y Acompañamiento Pedagógico dirigidos a docentes de aula desarrollados en modalidad presencial y virtual en el IESPP CREA. Ha participado en investigaciones sobre el desarrollo y aprendizaje en la infancia y adolescencia. ORCID ID: 0000-0002-8378-9013. Email: liz.ysla@gmail.com

Vicenta Ávila Clemente

Profesora titular de la Universidad de Valencia-España. Pertenece al grupo de investigación la ERI Lectura de la Universidad de Valencia. Ha desarrollado su investigación en el área de la discapacidad. En los últimos años está participando en proyectos competitivos relacionados con la lectura y comprensión de textos en distintas poblaciones. ORCID ID: 0000-0002-2762-2964. Email: Vicenta.Avila@uv.es