

Evaluación estandarizada de los aprendizajes: una revisión sistemática de la literatura

Standardized Learning Assessment: A Systematic Literature Review

Jorge Gustavo Gutiérrez Benítez^a


Luis Alan Acuña Gamboa^b


Recibido: 21 de mayo de 2021

Aceptado: 11 de octubre de 2021

Resumen: En el presente artículo se expone una revisión de la literatura que da cuenta del panorama de la evaluación de los aprendizajes bajo el enfoque estandarizado y con mediciones particulares de la psicometría. Con base en un sistema de búsqueda booleano y parametrizado por categorías analíticas, se conformó un corpus de estudio de 68 documentos recuperados de las bases de datos Dialnet, Scielo, Redalyc, Latindex, Elsevier, etcétera, los cuales fueron seleccionados bajo rigurosos criterios de inclusión y exclusión diseñados expresamente para esta revisión. Se empleó el análisis interno de contenido para la revisión del corpus de estudio. Entre los hallazgos más relevantes se aprecia un vacío entre los *softwares* educativos con funciones evaluativas y la capacidad de estos para realizar análisis psicométricos.

Palabras clave: evaluación de estudiantes; aprendizaje; psicometría; *software* educativo; pruebas.

^aEstudiante del Doctorado en Innovación en Tecnología Educativa de la Universidad Autónoma de Querétaro. Técnico Académico de Tiempo Completo de la Universidad Autónoma de Baja California, México.  jorgegustavo.gutierrez@gmail.com || ORCID: <https://orcid.org/0000-0003-3392-6398>.

^bDoctor en Estudios Regionales por la Universidad Autónoma de Chiapas. Docente Investigador de la Universidad Autónoma de Chiapas. Docente Investigador del Núcleo Básico del Doctorado en Innovación en Tecnología Educativa, Universidad Autónoma de Querétaro.  acugam@gmail.com || ORCID: <https://orcid.org/0000-0002-8609-4786>.

Abstract: This article presents a literature review on the panorama of learning assessment under the standardized approach doing a special emphasis in psychometric measurements. Based on a parameterized search system by analytical categories, the study corpus was form by 68 documents retrieved from databases like Dialnet, Scielo, Redalyc, Latindex, Elsevier, etc., which were selected under rigorous inclusion and exclusion criteria designed specifically for this review. Internal content analysis was used for the study corpus review. Among the most relevant findings, it can be seen a gap between educational software with evaluative functions and their capacity to perform psychometric analysis.

Keywords: student evaluation; learning; psychometry; educational software; tests.

Introducción

La evaluación estandarizada del aprendizaje, llamada también evaluación de gran escala, es una forma de evaluación sistematizada del aprendizaje, caracterizada por seguir un proceso riguroso con marcos referenciales teóricos y metodológicos con los cuales se miden rasgos observables en la población objeto, estableciendo precisiones específicas, controles logísticos y administrativos (Backhoff, 2018; Fernández, Alcaraz y Sola, 2017; Tristán y Pedraza, 2017). De igual manera, se identifica por la sistematización de los instrumentos o técnicas con las que se recopila, analiza e interpreta la información, de forma tal que se utilicen los mismos instrumentos durante todo el proceso (Jornet, 2017).

Las pruebas estandarizadas han surgido como una posible respuesta o herramienta para la mejora de los procesos educativos en las instituciones, en particular, con el fin de contar con instrumentos de evaluación que sean válidos y confiables, aportando información con la que instituciones educativas puedan tomar decisiones y emprender acciones para mejorar la calidad de sus procesos y, con ello, el aprendizaje de sus estudiantes (Gómez, 2004; Tiramonti, 2014). Por esto, las pruebas estandarizadas instauran parámetros de desempeño imprescindibles para establecer objetivos educativos, o se utilizan como índices para predecir el desempeño de estudiantes, entre otros usos.

Los retos que significa la evaluación de los aprendizajes en la educación superior, tanto como actividad que responde a las demandas y necesidades formativas del siglo XXI, así como mecanismo de análisis e intervención en el campo de la investigación educativa, hacen del tema un campo de estudio apremiante. De esta manera, el presente estudio da cuenta del desarrollo del conocimiento científico que se ha realizado en los últimos años sobre la evaluación estandarizada de los aprendizajes, en términos de aplicación de la psicometría como método de medición y aseguramiento de la calidad, así como la inclusión de la tecnología aplicada a los procesos educativos y de evaluación.

La estructura del documento se relaciona con tres grupos de análisis que corresponden directamente con el enfoque de la evaluación estandarizada de los aprendizajes. Por su parte, en el apartado de metodología se describen cuáles fueron los criterios de inclusión y exclusión de las fuentes consultadas para la investigación, dando así pertinencia a la integración y conformación del corpus. De igual forma, en este apartado se explica el procedimiento que se implementó para definir las distintas clasificaciones de interés en la literatura revisada. A guisa de cierre del artículo, se realiza un análisis de los vacíos de conocimiento existentes en relación con la aplicación de tecnología en la elaboración y puesta en práctica de pruebas estandarizadas, haciendo un análisis particular en los aspectos involucrados para determinar la calidad psicométrica de las mismas.

1 Metodología

La investigación se realizó a partir de un análisis profundo de carácter interno sobre el corpus diseñado expreso para este propósito (Jiménez, 2004; López, 2002; Urrutia y Bonfill, 2010). De esta manera, se entiende por análisis de contenido al conjunto de procedimientos que permiten realizar clasificaciones e inferencias, tanto válidas como reproducibles, de uno o varios textos que se relacionan entre sí por abordar un mismo tema de investigación; así mismo, hace hincapié en la exposición de los vacíos o lo no dicho en la materia (Aigner, 1999; Ulloa, 2015). El análisis realizado es de corte cualitativo por la pretensión de profundizar e interpretar los elementos más importantes de los documentos rectores en este trabajo, logrando así estimar los avances, retrocesos y retos en el campo de la evaluación estandarizada de los aprendizajes en la educación superior.

El corpus de estudio se conformó por 68 documentos recuperados de bases de datos de alto impacto, tales como Dialnet, Scielo, Redalyc, Latindex, DOAJ, Elsevier, bajo un sistema de búsqueda y selección de tipo booleano y parametrizado por categorías analíticas, con énfasis de indagación en títulos, resúmenes y palabras clave de los textos revisados (Figura 1), considerando los siguientes criterios de inclusión y exclusión: para la inclusión, 1) los documentos seleccionados debieron publicarse entre 1995 y 2020; 2) que en el resumen, introducción o prefacio se enunciara al menos una de las categorías analíticas del estudio (evaluación, psicometría y *software*); 3) que en todos los documentos seleccionados con los criterios 1 y 2, dentro del extenso se hiciera énfasis en la evaluación estandarizada de los aprendizajes; se excluyeron los documentos que 1) a pesar de cumplir con los criterios de inclusión, se encontraran en bases de datos, librerías o tiendas electrónicas con costo para su adquisición, o 2) que se encontraran en repositorios con acceso institucional único y exclusivo.

Tabla 1. Estructura del *corpus* de investigación

Tipo de publicación	Cantidad revisada	Tiempo comprendido
Artículo en revista indexada	54	1997-2019
Artículo en eventos académicos	5	1997-2019
Libros	4	1999-2014
Capítulo de libro	3	2006-2010
Tesis de grado	2	2000-2006

Fuente: elaboración propia.

Dicha revisión de la literatura se desarrolló tomando como líneas de análisis las categorías, variables e indicadores diseñados exprofeso para esta investigación (Figura 1). A partir de la técnica del fichaje se logró operativizar los ejes de macro (categorías), meso (variables) y micro (indicadores) análisis con los que se recuperaron los puntos de encuentro, desencuentro, tensiones y vacíos de conocimiento en este campo de investigación educativa, y que dan sustento a la caracterización de las prácticas de la evaluación de los aprendizajes, las metodologías que se utilizan para una evaluación de calidad, los principales atributos con los que se puede medir la calidad de una prueba estandarizada, así como las teorías sobre las que recaen todas las prácticas, técnicas o instrumentos que se emplean en la consecución de la evaluación estandarizada del aprendizaje.

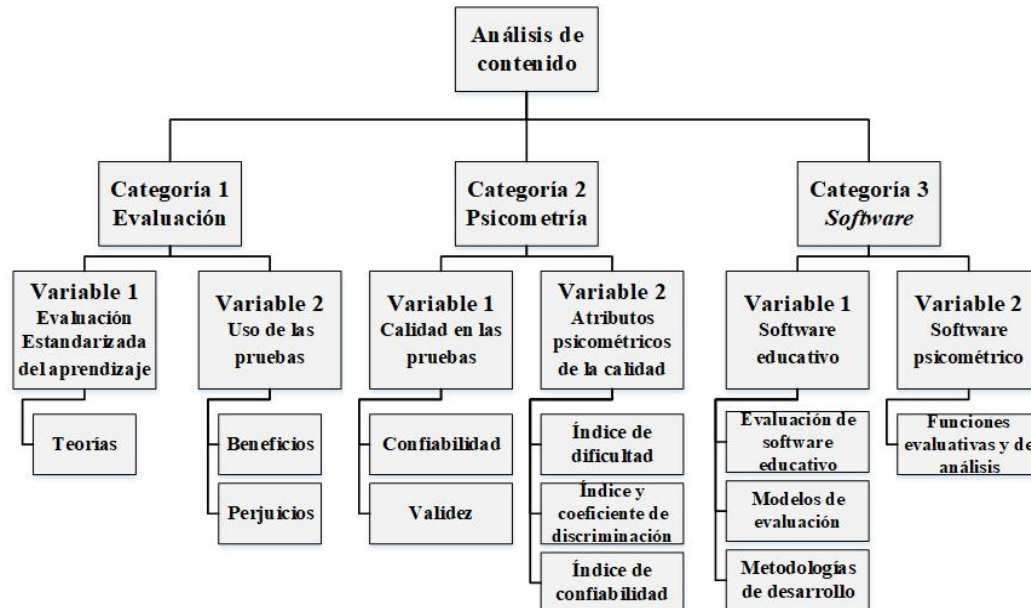


Figura 1. Categorías, variables e indicadores de análisis en la investigación

Fuente: elaboración propia.

Con base en la estructura de análisis de investigación señalada en la Figura 1 se procedió a la revisión sistemática de cada uno de los elementos que componen el corpus de la investigación, efectuando así un análisis cualitativo de estos documentos enfocando los esfuerzos en encontrar elementos relacionados con los indicadores definidos.

2 Ejes de la evaluación de los aprendizajes: la revisión

2.1 Evaluación del aprendizaje y pruebas estandarizadas

Las investigaciones que aquí se presentan se alinean al estudio del qué es y cómo se comprende la evaluación del aprendizaje, con un particular énfasis en las evaluaciones a gran escala, también conocidas como evaluaciones estandarizadas, en cuanto metodologías en el diseño o desarrollo de estas, y las teorías que fundamentan las técnicas o instrumentos con los cuales se elaboran este tipo de pruebas, cerrando con los usos y/o beneficios que se han logrado en la aplicación de éstas.

Definir la evaluación del aprendizaje obliga a la reflexión sobre las diferentes miradas que posicionan su objeto de estudio en este campo; esto con la finalidad de conceptualizarla en su justa medida, aunado al hecho que ésta es una temática de interés actual para las instituciones de educación superior por la creciente exigencia sobre la calidad en la formación de las nuevas generaciones de profesionales.

La evaluación del aprendizaje puede clasificarse desde diferentes líneas y con base en una diversidad amplia de criterios; por ejemplo, el momento en que se evalúa (inicial o diagnóstica); por el actor que la realiza (autoevaluación, heteroevaluación, coevaluación, etcétera); por la forma en que ésta se realiza (formativa, dinámica, estática, sumativa), o por el enfoque, como lo es la evaluación auténtica o alternativa centrada en los procesos (Izquierdo, 2008; Mora, 2004; Zúñiga, Solar, Lagos, Báez y Herrera, 2014). En este sentido, la evaluación de los aprendizajes funge como base fundamental para la observación de los aciertos y errores en las estrategias empleadas con el fin de lograr la adquisición de conocimientos en los estudiantes, y con ello tomar decisiones en pro de su formación educativa. Bogantes (2015) comenta que la evaluación del aprendizaje es el ejercicio educativo o formativo que dictamina qué, cómo, por qué y cuándo se debe enseñar; es decir, se asocia la evaluación del aprendizaje con las decisiones que se hayan tomado o se tomen durante los procesos de enseñanza y aprendizaje.

En este tenor, la evaluación es un proceso continuo que se realiza prácticamente desde que se tiene conciencia; por ello, Mendivil (2012) asevera que no surge en y para el ámbito pedagógico, ya que su génesis se alinea a la vida cotidiana, puesto

que en todo momento efectuamos una evaluación del comportamiento, la expresión, el rendimiento. Por otro lado, en el ámbito de la educación superior se observa una tendencia por describir la evaluación como el producto de una serie de interacciones entre diferentes tipos de personas en un espacio y tiempo determinados, interacción que particularmente ofrece un marco referencial sobre el cual los docentes orientan las prácticas de evaluación (Contreras, 2010).

A pesar de la relevancia que enmarca al campo, en la actualidad existe una confusión o mala interpretación del significado real de la evaluación del aprendizaje, que llega a reducirla a la mera acción de calificar; a este respecto, Alcaraz (2015) explica que normalmente se dice que se evalúa para ver si el alumno está aprendiendo o no, pero esto relega a último término la principal función de evaluar, que no concluye con la comprobación del aprendizaje, sino más bien con asegurar las condiciones para que ese aprendizaje se logre.

En general, se observan definiciones de evaluación del aprendizaje que se pueden clasificar o resumir en una serie de fines particulares. Uno de ellos es la evaluación como mejora, que para ciertos autores se comprende por dos tipos de evaluación: la evaluación formativa y la evaluación formadora (Umaña, Calvo y Salas, 2017); este tipo de evaluación permite transformar la práctica del docente y el aprendizaje del estudiante. Otro es la evaluación como rendimiento de cuentas del docente o de la institución, que se observa cuando los docentes demuestran que realizan correctamente su trabajo y que logran los objetivos educativos planteados por la institución (éste es uno de los usos que se les da a las pruebas estandarizadas, también conocidas como pruebas de logros). Otro más es la evaluación como rendimiento de cuentas del alumno, donde el estudiante demuestra, en la mayoría de los casos, el aprendizaje alcanzado durante un determinado tiempo (Hidalgo y Murillo, 2017). En este sentido, se identifica a la evaluación como una herramienta para la rendición de cuentas; no se limita a informar sobre los aciertos y desaciertos de un plan o programa de estudios o del desempeño profesional, sino también implica, de forma prácticamente obligada, la retroalimentación para el mejoramiento académico, tanto del personal docente como de la población estudiantil y, desde luego, de la institución educativa en su conjunto (Mora, 2004).

Por otro lado, la evaluación estandarizada de los aprendizajes según lo explican Tristán y Pedraza (2017), es uno de los instrumentos de prueba o medición más utilizados, que poseen un amplio desarrollo técnico y metodológico que los dotan de la capacidad de medir rasgos latentes u observables en la población con alto grado de precisión.

En relación con esta característica metodológica se observó también que la estandarización se entiende como un proceso de sistematización de todos aquellos

elementos que están asociados a una acción de recogida e interpretación de información, de tal forma que se utilicen los mismos instrumentos o técnicas tanto para recopilar como analizar e interpretar la información (Jornet, 2017). Es importante notar que otros autores apuntan como características particulares de este tipo de pruebas los marcos de referencia teóricos y metodológicos rigurosos (Backhoff, 2018; Fernández et al., 2017; Gómez, 2004; Martínez, 2001; Popham, 1999), resaltando el hecho de que se les asocia con fines de evaluación válidos y confiables. En las concepciones que se tienen de evaluación estandarizada se hace un especial énfasis en el elemento base de que son sistematizadas y que emplean métodos o instrumentos muy rigurosos para la recopilación y análisis de información. Se sustentan en análisis con marcos referenciales teóricos con los cuales se pueden efectuar mediciones que dan como resultado valoraciones cuantificables de atributos asociados a la calidad de la prueba, como son la validez y la confiabilidad.

2.1.1 Usos de las pruebas estandarizadas

Un tema asociado directamente a las pruebas de evaluación del aprendizaje es el uso que se le da a éstas. Uno de los propósitos o intenciones de evaluar es la obtención de información específica con base en la cual se puedan efectuar acciones o decisiones educativas. En este sentido, los usos que se le otorgan a las pruebas estandarizadas normalmente se encuentran asociados a pruebas de gran escala o de objetivos, donde, por citar un ejemplo, se encuentran las pruebas de admisión a las universidades y aquellas para medir el desempeño de los estudiantes de un país. Inicialmente se revisa la percepción general de los usos de este tipo de pruebas, y posteriormente se analizan las diferencias o aspectos negativos que se observan sobre ellas.

En el caso particular de México, este tipo de pruebas no ha sido una práctica muy frecuente; la Secretaría de Educación Pública empezó a utilizarlas en 1972 para decidir la admisión de alumnos en secundaria (Martínez, 2009). Sin embargo, la situación no tuvo otro avance sobresaliente hasta 1994, cuando se impulsó la realización de pruebas a gran escala en la educación básica mediante el proyecto denominado Estudio de Evaluación de la Educación Primaria.

Este año también fue importante para la evaluación en México, ya que con el ingreso a la Organización para la Cooperación y el Desarrollo Económicos (OCDE) se aumentaron los esfuerzos por integrarse a la vida económica y política internacional, lo cual incluía las evaluaciones educativas a gran escala. Esto produjo en consecuencia la creación del Centro Nacional para la Evaluación de la Educación Superior (CENEVAL), y con ello se extendió en el país este tipo de evaluaciones a

gran escala o estandarizadas. En 1996 se realizó un trabajo de estándares curriculares, produciendo evaluaciones en relación con ellos, lo que se conoció como Pruebas de Estándares Nacionales, aplicadas por primera vez en 1998 (Martínez, 2001).

Como se mencionó al inicio de este trabajo, una variable de interés es observar cuáles son los usos que se le dan a las pruebas estandarizadas. En tal sentido, sobresalen apreciaciones que asocian este tipo de pruebas con el rendimiento escolar, la calidad educativa en las instituciones e, incluso, con dar razones de la educación de un país (Gómez, 2004; Tiramonti, 2014).

Una característica de este tipo de pruebas es el exhaustivo control y método con el cual se realizan, y a este respecto, Jornet (2017) destaca la relevancia de utilizarlas al decir que el fin perseguido es lograr un sistema de acercamiento a la realidad, asegurando que la variación en los resultados sea atribuible al sujeto evaluado o a ciertos factores de intervención, pero no a la calidad técnica del instrumento o al proceso de elaboración del mismo.

Por su parte, algunos autores (Fernández et al., 2017) consideran que la importancia de usar estas pruebas radica en el hecho de que valoran el aprendizaje de los estudiantes de una forma masiva y lo atribuyen a los efectos del sistema educativo al que pertenecen, con lo que se permite proporcionar información sobre las fortalezas o debilidades de éste. Lo anterior se explica al decir que el sistema educativo produce aprendizajes en los estudiantes, y la aplicación de pruebas estandarizadas es la medición de los resultados de dichos aprendizajes, con lo cual se puede valorar la calidad del propio sistema educativo, y con ello señalar qué cambios deben realizarse para mejorar.

Autores como Shepard (2006) y Ravela (2010) explican que un beneficio de este tipo de pruebas es que permiten tomar conciencia acerca de la importancia de ciertos temas y capacidades que los estudiantes deben adquirir y desarrollar, y que regularmente no se encuentran dentro del marco conceptual sobre el cual los docentes organizan su enseñanza, situación que estimula la creación de nuevos marcos conceptuales sobre la didáctica e instrumentos o formas con las que se evalúa a los estudiantes. De esta forma, tales pruebas, cuando son alineadas con objetivos de aprendizaje, pueden ser muy útiles, ya que proporcionan la base para procesos de retroalimentación que el docente puede emplear para identificar fortalezas y debilidades curriculares, y desde luego verificar el logro alcanzado de los estudiantes, pero a nivel individual.

Resulta interesante observar cómo el planteamiento principal de estos autores sobre la virtud de la prueba estandarizada es tratar de contar con un instrumento

correctamente diseñado que minimice el error de medición, es decir, que las diferencias en los resultados al ejecutar determinada prueba se deben a las características propias del individuo y no a un mal diseño o estructura de la prueba.

Sin embargo, se esbozan también ciertos aspectos que tienen que ver con el uso de estas pruebas y que no son del todo favorables o acertados para evaluar el aprendizaje de los estudiantes, o más bien la calidad educativa de la institución.

2.1.2 Contraparte de los usos de las pruebas estandarizadas

Como se mencionó anteriormente, hay autores que manifiestan las ventajas y usos de este tipo de pruebas (Jornet, 2017; Ravela, 2010; Tiramonti, 2014), pero, así como han surgido usos y ventajas en la aplicación de pruebas estandarizadas, también existe una contraparte que expresan otros investigadores sobre las desventajas de utilizarlas. En este sentido, Popham (1999), haciendo referencia a las pruebas para medir la calidad educativa de una institución, considera:

Emplear pruebas estandarizadas de logros para averiguar la calidad educativa es como medir la temperatura con una cuchara. Las cucharas tienen la misión de medir cosas diferentes que el calor o el frío. Las pruebas estandarizadas de logros tienen la misión de medir algo distinto que cuán buena o cuán mala es una escuela. Las pruebas estandarizadas de logros deberían usarse para hacer las interpretaciones comparativas que se supone deben suministrar. No deberían ser usadas para evaluar la calidad educativa. (p. 4)

A esta línea de intelección se agregan otros autores que con el paso de los años han observado un uso desvirtuado de este tipo de pruebas, y cómo es que se ha llegado a conclusiones erróneas sobre la calidad educativa de una institución por utilizar resultados que midieron todo menos la calidad de la institución. Por ejemplo, Gomez (2004) explica que el conocimiento que se tiene sobre el rendimiento de los estudiantes se limita al desempeño obtenido en las pruebas estandarizadas; hay un desconocimiento de otras habilidades, actitudes y competencias que son implícitas a cualquier área de conocimiento. Y por esta razón las conclusiones que se emiten sobre la calidad de un sistema educativo son incompletas.

Siguiendo esta misma crítica, se agregan elementos como la mala práctica de preparar a los estudiantes exclusivamente para lograr un buen rendimiento en pruebas estandarizadas, debido a la presión administrativa por obtener un buen *ranking* y con ello acceder a ciertos apoyos o bien lograr un estatus social. De acuerdo con esta situación, se hace notar cómo los estudiantes son orientados, enseñados o educados

para responder correctamente a estas pruebas, en lugar de enseñarlos a pensar por sí mismos y convertirse en aprendices creativos (Moreno, 2016).

La revisión ha arrojado otros resultados sobre el mal uso de las pruebas, las investigaciones ponen de manifiesto situaciones desfavorables que trascienden la cuestión técnica de la prueba y, con esto, se hace referencia a la manipulación de la información por parte de las instituciones u organismos encargados de aplicar estas pruebas.

En relación con lo anterior, se encontró que ciertas instituciones manejan la información de manera secreta, cuando en realidad debería ser difundida; desde luego, tomando en cuenta controles de integridad y confidencialidad de la información de los participantes directos de la prueba, ya que compartir dichos resultados es clave para tomar acciones de mejora hacia el interior e incluso el exterior de la institución. Sin embargo, este tipo de prácticas en ocasiones son realizadas por motivos políticos (Martínez, 2001).

Un punto común que expresan estos autores es que se suele tomar decisiones con información o resultados de pruebas que no fueron diseñadas para ello. Además, se hace referencia a que las instituciones manipulan los resultados de una manera hermética, con lo que se entorpece la difusión de éstos, difusión que es elemental para lograr la mejora de la misma institución.

Es interesante cómo lo anterior indirectamente dice algo sobre la concepción de la calidad técnica de las pruebas, ya que, aunque se encuentran opiniones desfavorables sobre las pruebas estandarizadas, la mayoría de dichas opiniones están enfocadas en el uso que se da a los resultados obtenidos con las pruebas, y no tanto a un mal diseño de éstas, a los métodos o técnicas de recopilación y análisis de resultados.

2.1.3 Teorías y fundamentos de las pruebas

El diseño de una prueba debe ir acompañado de un sustento teórico que sirva de referencia para la aplicación de una o más estrategias pedagógicas y, a su vez, entender qué influencia o corriente cognitiva ha modelado su elaboración. En este sentido, es importante observar cuáles han sido las principales teorías sobre las que gira la evaluación del aprendizaje y el diseño de pruebas.

Una primera aproximación hace referencia al positivismo del siglo XIX, con los trabajos de Mill en 1822 y Comte en 1842. Binet (citado en Tristán y Pedraza, 2017) explica que:

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. (p. 18).

Por su parte Martínez (2001), en relación con el principio de las teorías de medición (término intrínseco a la evaluación), indica que surge siglos más atrás. Explica que las bases de la teoría de la medición fueron puestas por los trabajos de Laplace en 1796, que sentaron las bases para la teoría de probabilidades, y los de Gauss en 1798, con los fundamentos de la teoría de números. El interés por aplicar estas teorías de medición en contextos educativos surge con mayor auge en países como Alemania, Inglaterra, los Estados Unidos y, en menor medida, Francia y las regiones francófonas de Suiza y Bélgica.

Más adelante, a principios del siglo XX, Binet construyó el primer test estandarizado de inteligencia, teniendo como referencia los trabajos de Pearson y Spearman (Aiken, 2003), y como resultado permitieron la creación de la que después sería conocida como Teoría Clásica de los Test (TCT), enfoque teórico que tendría su mayor impulso y difusión en los años 50 (Muñiz, 2010).

El avance en el campo de la medición y elaboración de pruebas gracias a la TCT derivó en el surgimiento de otras teorías, como la Teoría de la Generalización desarrollada a comienzos de los años 70 (Quero, 2010) y la Teoría de Respuesta al Ítem (TRI), que se considera tiene sus inicios con los trabajos del modelo de Rasch, Lord y Novick en los años 60, pero que logró su mayor difusión con los trabajos de Lord a comienzos de los años 80 (Attorresi, Lozzia, Abal, Galibert y Aguerri, 2009).

Ambas teorías son propuestas para mejorar aquellos aspectos susceptibles de error y ausencia de medición en la TCT. En el caso particular de la TRI, se busca un fundamento probabilístico al problema de medir constructos latentes (particularmente los no observables), en el cual se considera al ítem como la unidad básica de medición, con lo que se puede pronosticar cómo responderá un sustentante a un ítem en particular (Cortada, 2004). Lo anterior muestra una de las ventajas observadas de la TRI en relación a la TCT, y tiene que ver con el sesgo, ya que aquella proporciona un marco de referencia unificado para conceptualizar los sesgos a nivel del ítem.

Actualmente se han desarrollado diversos trabajos basados en la TRI (Baladrón, Sánchez, Romeo, Curbelo, Villacampa y Jiménez, 2018; Ferreyra y Backhoff, 2016; Santelices y Valenzuela, 2015;), los cuáles muestran que esta teoría es aplicable en diferentes áreas del conocimiento.

2.2 Psicometría

2.2.1 Calidad de las pruebas

Es de suma importancia contar con instrumentos de evaluación que estén correctamente diseñados, con el fin de que los resultados que se obtengan sean válidos y confiables, para a partir de ellos hacer inferencias, emitir juicios y una toma de decisiones acertada.

En la investigación, una categoría de estudio es la psicometría, para la que se han definido dos indicadores de interés particular asociados a atributos psicométricos, a saber, la validez y la confiabilidad de una prueba. De igual manera, otro indicador asociado son los cálculos específicos que se realizan para determinar los atributos antes mencionados.

Haciendo una revisión inicial sobre la confiabilidad se encuentra que este criterio se asocia con los errores de medición (Argibay, 2006); el autor explica que en particular se señalan dos tipos de errores: los aleatorios que, dada su naturaleza, no hay una capacidad de control o predicción de los mismos, y los sistemáticos, que pueden ser controlados y ser sujetos de modificarse mediante alguna alteración en el mismo sistema. Para efectos de la calidad de una prueba, el error sobre el que hay interés es el aleatorio. Es así como en toda medida el valor obtenido es compuesto por el valor verdadero y los errores en la medición, de forma tal que un instrumento será más confiable en razón de maximizar el valor verdadero.

En términos académicos se ha encontrado que el concepto de confiabilidad se explica, por ejemplo, si, en ausencia de cualquier cambio de manera permanente en una persona, las calificaciones de una prueba varían en gran medida con el tiempo o en diferentes situaciones, es probable que la prueba no sea confiable, y por tal razón no pueda ser utilizada para explicar o predecir el comportamiento de los sustentantes (Árraga y Sánchez, 2012).

Como se observa, las investigaciones manifiestan la confiabilidad como un atributo elemental para considerar la calidad de una prueba. Sin embargo, la confiabilidad debe ir acompañada de otro atributo, la validez.

Una prueba puede ser confiable, pero si no es válida, los juicios emitidos a partir de los resultados serán erróneos o insuficientes para tomar decisiones.

Esto sucede con el uso que se hace de las pruebas; páginas atrás se incluyó un fragmento en el que Popham (1999) mencionaba que las pruebas estandarizadas eran similares a medir la temperatura con una cuchara, es decir, que el instrumento no

era válido para lo que se deseaba evaluar. Lo anterior significa que de nada sirve tener un instrumento confiable si éste no es el adecuado para la medición que quiere realizarse.

Continuando con esta idea, en la conceptualización de validez varios autores (Aliaga, 2006; Árraga y Sánchez, 2012) concuerdan en que se trata de un juicio evaluativo, en el cual existe una evidencia empírica y supuestos teóricos que dan un respaldo a la suficiencia y lo apropiado de las interpretaciones y acciones en base a los puntajes de las pruebas, que no se limita a los reactivos de la prueba, sino que también incluye la forma en que los sustentantes responden y el contexto en que se desarrolla dicha evaluación. Lo anterior se puede resumir como el grado en que un instrumento de evaluación realmente mide aquello para lo cual fue diseñado. Existen diferentes tipos de evidencia de validez, por ejemplo, la asociada con el contenido, la que se relaciona con el criterio y la referida al constructo.

Sin embargo, Gregory (en Árraga y Sánchez, 2012) señala que algunos autores e investigadores en psicometría, como Cronbach, Guion y Messick, consideran a la validez de constructo como el elemento central para todos los tipos de evidencia de validez, dejando los restantes tipos solo como apoyo.

2.2.2 Análisis psicométrico y medición de la calidad

En apartados anteriores se analizó cómo surgieron las primeras teorías que sentaron las bases para el desarrollo de pruebas. En el caso particular de la TRI se han efectuado investigaciones para determinar una serie de atributos con los cuales se puede observar la validez y confiabilidad de una prueba. En esta investigación, una variable de estudio definida es la calidad de las pruebas mediante el uso de la psicometría. Para esta variable se definieron tres indicadores de interés especial, que a su vez son los atributos con los que se observa la calidad técnica de una prueba basada en la TRI: el índice de dificultad del ítem, el índice de discriminación y el coeficiente de discriminación.

El índice de dificultad del ítem se define como la proporción de una muestra o población que responde acertadamente un ítem o pregunta en una prueba (Medina, Ramírez y Miranda, 2019). Croker y Algina (en Backhoff, Larrazolo y Rosas, 2000) mencionan que, usualmente, a esta proporción se le denota con una p , la cual indica la dificultad del ítem. El cálculo de este atributo se realiza mediante la división del número de personas que contestó acertadamente el ítem entre el número total de personas que lo contestaron.

En cuanto a los valores posibles de este atributo, van desde cero hasta uno; Wood explica (en Backhoff et al., 2000) que “a mayor dificultad del ítem, menor será su

índice” (p. 14). Lo anterior quiere decir que entre más cercano a uno se encuentre el valor de este índice, el ítem es más fácil de responder, y viceversa.

En relación con el índice de discriminación de la prueba, la revisión encuentra que los autores (Backhoff et al., 2000; Medina et al., 2019) han definido este atributo con una analogía simple: quien haya obtenido una mejor puntuación en todo el examen deberá tener mayores probabilidades de contestar correctamente un ítem; así pues, la discriminación es la cualidad que tiene un ítem para separar a los estudiantes con mejores puntuaciones de aquellos con menor puntuación final en la prueba.

Existen diferentes métodos para obtener la discriminación; uno de ellos consiste en separar a la población de los sustentantes en dos grupos, 50% con puntajes superiores a la media y 50% con puntajes inferiores.

Otro método encontrado en la revisión de este apartado es tomando percentiles de la población, en lugar de considerar a todos los sustentantes. En este sentido, Backhoff et al. (2000) sólo consideran a 54% de ellos: 27% de las puntuaciones más altas en el test y 27% de las puntuaciones más bajas en el test. Mientras que Medina et al. (2019) hacen el cálculo con 25% de la población tanto en grupos altos como bajos.

Si bien existe una diferencia entre los autores referente a los percentiles al momento de calcular el índice, donde hay concordancia es en la ventaja que presenta emplear este método, ya que para ambos la ventaja consiste en reducir la probabilidad de subestimar el nivel de discriminación de los ítems, precisamente por incluir sólo a aquellos sustentantes con mayor consistencia en su rendimiento. Para este atributo el rango de valores es de $[-1, 1]$; respecto a estos valores Ebel y Frisbie (en Backhoff et al., 2000) proponen una regla para clasificar la calidad del ítem en términos de índice de discriminación. Como mínimo este índice debe encontrarse por encima de .2 para considerarse regular, entre .3 y .39 se considera buena, y mayores a .39 son excelentes. Un índice por debajo de .2 significa que el ítem es defectuoso, necesita revisión profunda o bien desecharse.

Además del índice de discriminación, también existe el coeficiente de discriminación (Medina et al., 2019; Pérez, Acuña y Arratia, 2008), conocido como el punto de correlación biserial (r_{pbis}), atributo que permite medir más acertadamente la discriminación de un ítem. Los autores antes referidos definen este coeficiente como una medida de la consistencia de un ítem con toda la prueba en su conjunto, el cual refleja la correlación entre los puntajes de los sustentantes en un ítem en particular y sus puntajes en la prueba completa.

Este coeficiente permite observar la probabilidad de que un ítem sea contestado correctamente por aquellos estudiantes con mayor puntuación en la prueba; esto se deduce al presentar una correlación positiva y que se encuentre cercana a uno, es decir, que entre más cercana a uno sea la correlación, la probabilidad es más alta. En el caso contrario, si la correlación es negativa, significa que habrá una mayor tendencia a que los estudiantes con menor puntuación en la prueba acierten el ítem, lo que puede significar que el ítem es defectuoso. Por lo tanto, el rango de valores posibles para este atributo es de $[-1,1]$.

Referente a los atributos que permiten observar la confiabilidad de una prueba mediante una medición psicométrica, están el Índice de Confiabilidad de Kuder-Richardson (KR20) y el Alfa de Cronbach. Ambos índices permiten medir la consistencia interna (Reidl, 2013) de instrumentos o pruebas, además de ser los procedimientos más comunes para dicho propósito. Para que el índice KR20 sea de calidad o aceptable, debe ser mayor a 0.70, mientras que para el Alfa de Cronbach debe ser de 0.80.

Otra forma de obtener una valoración de la confiabilidad de una prueba es mediante el método test retest, el cual consiste en aplicar la prueba en diferentes momentos a la misma muestra de sustentantes, para observar las fluctuaciones o variaciones en los resultados de estos. Autores como Robins (citado en Ezpeleta, de la Osa, Domenech, Navarro y Losilla, 1997) y Serra y Peña (2006) explican que este método de medición de la confiabilidad es efectivo porque evalúa la estabilidad de la medida en el tiempo, a pesar del cambio. Además, normalmente se utiliza por la simplicidad y el bajo coste que implica.

La valoración de confiabilidad mediante el método test retest se hace con el cálculo del coeficiente de correlación intraclass (CCI), conocido también como índice de concordancia (Mandeville, 2005); los valores aceptados para este coeficiente van de 0 a 1, donde entre más cercano a uno significa un mayor grado de acuerdo, según la escala de Landish y Koch (Etchezahar, Prado-Gascó, Jaume y Brussino, 2014).

Como se ha analizado en esta revisión de las distintas investigaciones, la mayoría de los autores contemplan como elementos básicos de validez de las pruebas las características de dificultad y discriminación de los ítems, con sus respectivas variantes, como se veía en el caso de Backhoff et al. (2000) y Medina et al. (2019). Se observa también una concordancia en relación a que los atributos de confiabilidad que tienen uso en la evaluación de la calidad de una prueba son el índice de consistencia interna, conocido como KR20, y el Alfa de Cronbach, aunque también se utiliza el método del test retest mediante el cálculo del CCI. Cabe destacar que la mayoría de las investigaciones que aplican este método está asociada a estudios de carácter médico o de pruebas clínicas con pacientes.

2.3 Tecnología, evaluación y pruebas

En este grupo la revisión trata sobre aquellas investigaciones que abordan la inclusión de tecnologías de la información y la comunicación (TIC) en procesos de evaluación y evaluación estandarizada del aprendizaje. De igual manera, se señala qué teorías o metodologías se han desarrollado para diseñar *software* con fines educativos, y cómo se evalúa la calidad de estos. Por otra parte, también se analizan qué *software* existe, con la particularidad de tener como fin específico el análisis psicométrico.

2.3.1 Software educativo

Con la socialización del Internet y el acceso a computadoras (Castells, 2010), la incursión de plataformas tecnológicas en contextos educativos no tardó mucho en surgir, ya que la tecnología permitió acceder a información de formas distintas a lo tradicional, facilitando la consulta libre, así como generar representaciones virtuales o de contenido enriquecido de casi cualquier tema que se deseara. Todo esto encontró también un espacio de oportunidad para el ámbito educativo, con la creación de las primeras herramientas tecnológicas diseñadas específicamente con fines educativos.

En términos conceptuales, varias investigaciones tienen similitudes (Almaguel, Álvarez y Pernía, 2016; Cataldi, 2000; Vidal, Gómez y Ruiz, 2010) en la definición de *software* educativo; a grandes rasgos se les ubica como programas computacionales que facilitan la enseñanza y el aprendizaje; algunos los señalan como facilitadores del aprendizaje, o que sólo pueden recibir el término de *software* educativo cuando el programa incorpora una intencionalidad pedagógica, que normalmente está orientada a un objetivo de aprendizaje.

Como se observa, aunque las investigaciones antes mencionadas se encuentran registradas en un lapso de 15 años, concuerdan en que la característica principal es la inclusión de aspectos pedagógicos propios del proceso de enseñanza-aprendizaje, mediados o representados por un programa o aplicación. Es interesante cómo incluso se asocia directamente con la calidad, al comentar que es una herramienta a la que se le atribuye mejora en la calidad del proceso de enseñanza-aprendizaje (Gómez, 1997).

A este respecto, autores como Couturejuzón (2003), Reyes, Fernández y Duarte (2015) y Vidal et al. (2010) comentan que el *software* educativo es un producto tecnológico diseñado para apoyar procesos educativos, el cual permite al docente y al estudiante alcanzar de una mejor manera las metas de aprendizaje establecidas. Además de ser un medio de amplia difusión para representar y desarrollar

prácticamente cualquier contenido educativo, haciendo las veces de un libro, revista, video, dibujo, diagrama, etc., con su propia representación y formato digital. Lo que permite alcanzar una representación dinámica de todos los objetos, tales como figuras, diagramas o gráficas, así como de sonido para estimular o propiciar una mejor relación entre lo concreto y lo abstracto.

También se señala que para que un *software* sea considerado educativo debe cumplir con un principio básico, el cual consiste en integrar de forma casi homogénea tres ciencias, la computación, la pedagógica y la ciencia en cuestión para la cual se desarrolla el *software*.

Como se puede observar, las investigaciones antes citadas consideran como punto de especial interés la capacidad del *software* para representar casi cualquier contenido en formato digital, lo que facilita la exposición del estudiante a diferentes contenidos que pueden ser enriquecidos con objetos multimedia, simulaciones, videos, entre otros. Y que a su vez se asocia este tipo de elementos como un aporte a la calidad del proceso de enseñanza.

Lo anterior se concluye al considerar, por ejemplo, que este tipo de tecnologías facilitan el almacenamiento de información, el acceso a la misma y una mejor y más oportuna comunicación entre los diferentes actores que participan del proceso de enseñanza-aprendizaje, eliminando las barreras del espacio y el tiempo, permitiendo de esta manera a los estudiantes interactuar con objetos virtuales a los que tal vez en la realidad no pudieran acceder.

Si bien existe una variedad de *software* de características educativas, es de particular interés encontrar aquellos que utilicen estas tecnologías para evaluar el aprendizaje. A este respecto se encuentran investigaciones como la de López, Hernández y Farran (2011), quienes hablan acerca de una plataforma de evaluación automática con una metodología efectiva para la enseñanza-aprendizaje en programación de computadores. González (2006) habla sobre el desarrollo de un *software* educativo para el autoaprendizaje de bases de datos en asignaturas de nivel superior, con la capacidad de incluir un módulo para evaluar el aprendizaje. Estas investigaciones son ejemplos de *software* educativo con capacidad o funciones de evaluación dentro de los mismos. También existe software educativo con capacidades evaluativas que son del tipo de plataformas auto construibles, y que cuentan con una amplia difusión y uso por diferentes instituciones educativas, tales como *Blackboard*, *Google classroom* y *Moodle*.

Hay también *software* con fines específicos de evaluación del aprendizaje; por ejemplo, Tirado, Backhoff y Larrazolo (2016) hablan de cómo los exámenes más acreditados elaborados por el Educational Testing Service, como son el Test of

English as a Foreign Language (TOEFL iBT) y el Graduate Record Examination (GRE), son presentados o ejecutados en plataformas digitales. Referente a pruebas a gran escala o estandarizadas, se tiene por ejemplo la evaluación del Programme for International Student Assessment (PISA), que también ha migrado su presentación a formatos digitales. En el caso particular de México, desde 1993, el Examen de Habilidades y Conocimientos Básicos (EXHCOBA) se ha aplicado en formato digital como instrumento de selección para los aspirantes a instituciones universitarias de educación media superior y superior.

Siguiendo con la línea de *software* evaluativo en México, se tiene el Examen de Conocimientos Básicos, el cual posee características más complejas que sólo presentar preguntas de opción múltiple; está sustentado en tres soportes teóricos: la psicometría, teorías cognitivas y ciencias de la computación. Con lo cual se permite representar de manera digital reactivos estructurales constructivos, conocidos así porque la actividad a evaluar apela a un campo semántico del conocimiento, y no a un contenido puntual (Tirado et al., 2016).

El *software* educativo tiene fines pedagógicos en cualquier elemento del proceso de enseñanza-aprendizaje, ya sea como herramienta para la representación de conceptos, acceso ágil al almacenamiento y consulta de información, aumentar la comunicación e interacción, minimizar costos, riesgos y peligros mediante simuladores, y desde luego aprovechar estos recursos para evaluaciones, de carácter ordinario o estandarizadas.

2.3.2 Evaluación del software educativo

Este punto en particular ha sido investigado y puesto a discusión por varios autores, especialistas en educación, los cuales han generado como resultado una serie de normas, sugerencias o instrucciones con las cuales se puede evaluar un *software* en términos pedagógicos y técnicos (Galvis, 2000; Gómez, 1997; Marquès, 1998; Martínez y Sauleda, 1992; Navarro, 1999; Squires y McDougall, 1997).

Es comprensible que en la medida que se ha difundido el uso de estas tecnologías en los contextos educativos, ha sido proporcional el interés por valorar realmente cuál es el aporte significativo que tienen en relación al proceso de enseñanza aprendizaje. A este respecto Cataldi, Lage, Pessacq y García (1999) dicen: “La proliferación de estos materiales de apoyo educativo lleva consigo la necesidad de evaluar su calidad pedagógica y su pertinencia con el entorno en el cual se van a utilizar” (p. 187).

Es importante destacar que la evaluación de un *software* educativo depende de varios factores que determinan la misma, por ejemplo, el fin pedagógico deseado,

características de los usuarios finales del *software*, infraestructura tecnológica (velocidad de Internet, *hardware* e incluso otros *softwares*, por mencionar algunos) requerida para el correcto funcionamiento. Cova, Arrieta y Riveros (2008) en relación a lo anterior comentan que el *software* es valorado en razón de las funciones derivadas para el docente y de las interacciones que ocurran en el aula causadas directamente por el programa. Razón por la que en la evaluación deben incluirse las opiniones directas de los usuarios finales en el ambiente directo de uso, que en este caso se refiere al contexto de aprendizaje, además de las opiniones de los desarrolladores del programa (Cataldi, 2000).

Resulta notorio que los autores revisados consideran que un elemento crucial en la evaluación del *software* es la calidad pedagógica y la validez en el contexto final de aplicación, y que estos procesos de valoración deben considerar aspectos como el tipo de interacción e incluso las condiciones de infraestructura con las que se cuenta, y de esta forma determinar de una manera más acertada la calidad del *software* educativo.

2.3.3 Modelos de evaluación de software

Se ha revisado cómo es que para ciertos autores la evaluación del *software* educativo debe responder a las experiencias de aprendizaje que aporta y al enfoque de la enseñanza con el que se sustenta. Para esto y en pro de normalizar los procesos de evaluación de *software* se han desarrollado diferentes modelos de evaluación. Aunque existen muchos modelos y siguen surgiendo (esto en razón de los avances tecnológicos en las ciencias computacionales, las nuevas capacidades de procesamiento, representación de imágenes, realidad aumentada, etc.), la revisión se centra sólo en aquellos que han tenido mayor difusión tanto en la explicación de la evaluación como en la puesta en práctica.

A este respecto se mencionan los trabajos de Clarke, Pete y Naidoo (citados en Cova et al., 2008), que toman como eje central de evaluación tres dimensiones específicas: la pedagógica, la matriz de evaluación (con la cual se evalúan la conveniencia para el contexto de enseñanza, viabilidad, confiabilidad y validez), y la usabilidad del *software* (interface de interacción con el usuario).

También se encuentran las aportaciones de Poole (1999), que en su modelo establece lo que se conocen como Listas de control, que poseen una serie de indicadores de calidad tanto de la parte educativa como de la parte técnica del *software*. Estas listas de indicadores son desarrolladas considerando las necesidades individuales de la escuela y son especificadas por docentes, desarrolladores de *software* y asesores expertos en el área.

Otra metodología es la planteada por Marqués (citado en Cova et al., 2008); en ésta se presenta la Ficha de Catalogación y Evaluación Multimedia de los Programas Educativos, la cual recoge los rasgos fundamentales del *software* y las valoraciones sobre los aspectos funcionales, técnicos y pedagógicos.

Siguiendo con la revisión, se encuentra el modelo de evaluación de *software* educativo bajo enfoque sistémico de Díaz, Pérez, Mendoza y Grimán (2003). El modelo se basa en aplicar una serie de cuestionarios que permiten cuantificar la calidad en tres criterios principales: la usabilidad, la funcionalidad y la confiabilidad, que a su vez permiten medir de forma general la calidad del programa en tres categorías: baja, intermedia o avanzada. Al igual que el de Poole (1999), este modelo incluye la participación de especialistas en áreas de ciencias de la computación, además de docentes y estudiantes.

En una última revisión se encuentra un modelo propuesto por Straccia, Zanetti y Pollo (2019); éste toma como referencia los estándares de la Organización Internacional de Normalización (ISO, por sus siglas en inglés) ISO/IEC 9126:2001, ISO/IEC 14598:1999 e ISO/IEC 25000:2005 actuales, con los cuales se evalúa la calidad de un *software*, y se agregan aquellos aspectos distintivos del *software* educativo, como lo son atributos para medir el aprendizaje, la enseñanza, condiciones referidas a los destinatarios y contenido. Cada uno de estos atributos se mide con una serie de indicadores específicos.

Recopilando los aportes de los autores revisados en este apartado se puede notar que se tiene una fuerte tendencia a incluir a profesionales en el área de las ciencias computacionales en los modelos de evaluación del *software*, lo cual tiene sentido, puesto que hay muchos elementos programables que por su naturaleza de construcción pueden ir o no alineados a las funciones pedagógicas que la necesidad educativa demande, razón por la que sólo expertos en el área pueden determinar si son factibles o no de incluirse como requerimientos de funcionalidad o usabilidad en el *software* educativo.

También se observa que hay un común denominador en los autores al momento de clasificar la calidad del *software* mediante atributos observables, ya que los reducen a dos aspectos en concreto. El primero que tiene que ver con la parte educativa (llamada pedagógica en otros autores), y el segundo que comprende la parte tecnológica (llamada técnica por otros autores) que en algunos casos los autores la subdividen en dos o más atributos de calidad.

2.3.4 Ingeniería y metodologías para el desarrollo de software educativo

En relación con este punto, existen procesos propios de las ciencias de la ingeniería con los que se hacen aportes para el desarrollo de *software* educativo. Aquí se hace mención a unas cuantas propuestas para la mejora del desarrollo de *software* educativo producto de investigaciones correspondientes a la ingeniería de *software* y a las metodologías de desarrollo de *software*.

Con la incursión de la tecnología en los ambientes educativos, el desarrollo de *software* para estos fines representaba un reto. La percepción que se tenía del desarrollo de *software* educativo es captada por lo que dicen Cataldi, Lage, Pessacq y García (1999), quienes explican:

Uno de los problemas más importantes con los que se enfrentan los ingenieros en *software* y los programadores en el momento de desarrollar un *software* de aplicación, es la falta de marcos teóricos comunes que puedan ser usados por todas las personas que participan en el desarrollo del proyecto informático para aplicaciones generales. El problema se agrava cuando el desarrollo corresponde al ámbito educativo debido a la total inexistencia de marcos teóricos interdisciplinarios entre las dos áreas de trabajo. (p. 185-186)

Se puede notar cómo, aunque existía poco desarrollo o propuestas para desarrollar *software* con fines educativos, ya se hacía notar la necesidad y se vislumbraba un posible crecimiento de este tipo de desarrollos.

Las primeras aproximaciones o marcos de desarrollo de *software* educativo eran limitados; por ejemplo, Cataldi et al. (1999) proporcionan una explicación de los diferentes modelos o metodologías de desarrollo de *software* (señalados también como paradigmas de desarrollo) con la intención de que el diseñador del *software* empleara alguna de estas metodologías de acuerdo con el mejor acoplamiento o facilidad de uso en relación al tipo de proyecto (*software* educativo) a realizar, los recursos disponibles, etc. Se analiza cómo en línea con la teoría educativa y el currículo, se deberá adaptar alguno de los paradigmas del ciclo de vida, discriminando en cada etapa las actividades a realizar con la documentación, las técnicas y herramientas a utilizar según la necesidad pedagógica en cuestión.

Lo anterior deja ver que, en su mayoría, con lo que se contaba en esos años era con una mediana adaptación de la ingeniería de *software* a la producción educativa, ya que sólo se limitaba a dar una variedad de modelos o metodologías, y seleccionar aquella que se adecuara más a las necesidades del proyecto, pero no existía un marco referencial diseñado especialmente para desarrollar *software* de este giro.

Autores como Marquès (1998) proponían una metodología para desarrollar *software* educativo, pero ésta se concentraba en el aparato pedagógico como principal fuente; en otras palabras, era adecuar el *software* a un modelo pedagógico, mientras que, como se explicó anteriormente, la ingeniería de *software* de esos años era lo contrario, adecuar una necesidad educativa a un modelo de desarrollo de *software*.

Sin embargo, con el pasar de los años, el avance de las ciencias computacionales y la llegada de Internet (detonante en muchos otros aspectos tecnológicos y metodológicos), surgieron inevitablemente más investigaciones y trabajos en la búsqueda de crear un modelo o marco referencial que integrara un modelo para desarrollar *software* educativo de forma nativa.

Con el surgimiento de metodologías ágiles de desarrollo de *software*, como lo son Programación Extrema (XP, por sus siglas en inglés), CRYSTAL, SCRUM, surgieron propuestas metodológicas que integraban estos nuevos modelos en la ingeniería de *software* con fines educativos. Así surgieron investigaciones como la de Orjuela y Rojas (2008), quienes proponían un modelo basado en siete etapas propias o de carácter general para cualquier desarrollo de *software* con base en una metodología ágil, pero a su vez, cada una de estas etapas incluía ahora de forma nativa elementos propios del diseño educativo.

Así, por ejemplo, la etapa de planeación incluía un proceso de aspectos educativos; en la etapa de diseño se agregaban tres procesos más: diseño educativo, diseño de comunicación y diseño computacional, con los cuales se consideraban y observaban los elementos pedagógicos que el *software* requería en paralelo a las necesidades de usabilidad. Y lo mismo sucedía con las restantes cinco etapas del ciclo de desarrollo del *software*.

Con la creciente difusión de estas metodologías y la exposición social en masa al *software* gratuito en Internet con fines educativos, el acceso a sistemas o plataformas para desarrollar *software* personal y no de grado industrial, surgieron más propuestas metodológicas para desarrollar *software*, pero ahora con un enfoque particular en las características específicas del contexto educativo para el cual intervenía el desarrollo del *software*.

En apartados anteriores se observaba cómo autores que investigaron sobre evaluación de *software* educativo señalaban que un elemento importante era tener presente siempre las necesidades particulares de la institución, pero con los avances tecnológicos y metodológicos, las posibilidades fueron más allá, permitiendo desarrollar *software* para situaciones o momentos particulares del proceso de enseñanza-aprendizaje, mediante el desarrollo de metodología adaptables a contextos y situaciones particulares. En este sentido se recopilan aportaciones como

las de González (2006); García, Vite, Navarrate, García y Torres (2016); Madariaga, Rivero y Leya (2016), y Esterkin y Pons (2017).

2.3.5 Software psicométrico

Con relación a esta temática se rescatan aquellas investigaciones sobre *software* con fines psicométricos o con capacidades psicométricas. Se entiende por *software* con capacidades psicométricas aquel que puede realizar por lo menos un análisis psicométrico básico de una prueba, tales como los revisados en apartados anteriores, calcular el índice de dificultad de la prueba, índice de dificultad del ítem, índice y coeficiente de discriminación del ítem, índice de consistencia interna KR20, Alfa de Cronbach, entre otros.

Existen investigaciones que hacen uso de *software* especializado, tales como las de Cechova, Neubauer y Sedlacik (2014) y Thoe, Fook y Thah (2009), en las que se utiliza el *software* ITEMAN de la compañía *Assessment Systems Corporation*. Algunos otros trabajos de investigación (Backhoff et al., 2000) recurren al uso de *software* estadístico Statistical Package for the Social Sciences (SPSS) propiedad de IBM. También hay quienes implementan otro *software* (Marr, Gupchup y Anderson, 2012), conocido como R Studio, de distribución libre, con el cual se pueden efectuar análisis estadísticos. También existe el *software* Test Analysis Program (TAP) publicado por Brooks y Johanson (2003), el cual permite hacer análisis psicométrico; su versión gratuita está limitada a un cierto volumen de datos, sin embargo, permite obtener todos los atributos psicométricos mencionados en los apartados anteriores.

No obstante, después de revisar las investigaciones o aportaciones sobre este campo, se hace notar la ausencia de *software* con capacidades de análisis psicométricos que estén integrados a su vez en *software* con fines educativos, sobre todo en aquellos que están diseñados con fines evaluativos específicamente.

Esto señala que, primeramente, no todos los *softwares* poseen la capacidad de generar resultados en un formato o estructura base que sirva de insumo para *software* especializado para análisis psicométrico, como es el caso de TAP y de ITEMAN, y en cuanto a los *softwares* que se han revisado, no se integra en su diseño o funciones evaluativas la posibilidad de efectuar un análisis psicométrico en el mismo sistema; para esto recurren a un *software* de terceros, que en la mayoría de los casos que se analizaron necesitan un archivo fuente en un formato específico y particular. En la Tabla 2 se puede observar la lista de los *softwares* educativos y de análisis psicométricos que se revisaron en esta investigación.

Tabla 2. *Software* educativo y psicométrico revisado

<i>Software</i>	Tipo
Blackboard	De enseñanza y aprendizaje, con capacidad evaluativa y psicométrica (limitada)
Google Classroom	De enseñanza y aprendizaje, con capacidad evaluativa
Moodle	De enseñanza y aprendizaje, con capacidad evaluativa y psicométrica (limitada y requiere <i>plugins</i>)
EXHCOBA	Función específica de evaluar, diseñado con base en modelos y análisis psicométricos
EXCOBA	Función específica de evaluar, diseñado con base en modelos y análisis psicométricos
TAP	Función específica de realizar análisis psicométricos, requiere de archivos fuente en formatos específicos
ITEMAN	Función específica de realizar análisis psicométricos, requiere de archivos fuente en formatos específicos
R Studio	Puede realizar múltiples análisis estadísticos, entre ellos los análisis psicométricos, requiere archivos fuente en formatos específicos
STATA	Puede realizar múltiples análisis estadísticos, entre ellos los análisis psicométricos
SPSS	Puede realizar múltiples análisis estadísticos, entre ellos los análisis psicométricos

Fuente: elaboración propia.

3 Discusión y conclusiones

Con base en esta revisión resulta evidente que existen puntos de encuentro y desencuentro entre los autores. Por ejemplo, en cuanto a la evaluación del aprendizaje, se aprecia que en este campo se ha avanzado mucho y que ha sido un tema de constante investigación en las últimas décadas, que denota una influencia europea sobre las metodologías o técnicas que competen a esta área, así como las teorías sobre las que se sustentan.

Por otra parte, fue relevante observar, en cuanto a las pruebas estandarizadas, cómo un grupo de autores (Fernández et al., 2017; Jornet, 2017; Tiramonti, 2014) las defienden por su calidad técnica, por la utilización de métodos científicos, marcos de referencia teóricos y metodológicos rigurosos, mientras que algunos otros, aunque no las rechazan por completo, hacen notar o señalan puntualmente cuáles son sus defectos, tales como la tendencia a utilizarlas para medir aspectos de la educación para los cuales no fueron diseñadas las pruebas, la manipulación o manejo de estadísticos para acceder a *rankings* nacionales o internacionales, preparación excesiva del alumno para salir bien en dichas pruebas alejándolo así de un aprendizaje significativo, entre otros. Algunos autores (Gómez, 2004; Moreno, 2016; Popham, 1999) refieren que, en sí, no hay una mala concepción en el diseño

y la calidad técnica de las pruebas, sino más bien que normalmente se les ha utilizado para medir aspectos educativos que no son adecuados, mucho menos acertados. En ocasiones, por desgracia, los resultados de estas pruebas y sus malos usos tienen un alto impacto en las instituciones educativas.


Referente a las teorías y fundamentos de la prueba, se observó la tendencia moderna de aplicar, por ejemplo, la TRI por encima de la TCT o la Teoría de la Generabilidad; también se notó cómo la mayoría de los autores concuerda en que los atributos de validez y confiabilidad de la prueba pueden ser medidos con análisis de ciertos atributos psicométricos, como los índices de dificultad y de discriminación, pero resalta que se considera más efectivo el coeficiente de discriminación. Y que, en cuanto a la consistencia interna, la mayoría de los estudios considera de calidad los valores que se obtienen a partir del cálculo del KR20 y el Alfa de Cronbach.

En el último apartado (tecnología, evaluación y pruebas), se hizo notar la tendencia en los autores a resaltar la importancia que ha tenido el desarrollo tecnológico en la mejora de la calidad de la educación en general, y en particular de las capacidades para mejorar el proceso de enseñanza y aprendizaje (Almaguel et al., 2016; Cataldi, 2000; Couturejuzón 2003; Reyes et al., 2015; Vidal et al., 2010). Se dio cuenta del surgimiento de los *softwares* educativos y cómo estos en sus primeros años no tenían un marco de referencia para poder ser diseñados y construidos, tuvieron que pasar un par de años y un nuevo salto tecnológico para facilitar estos procesos, que a su vez también fueron impulsados por los mismos cambios en esquemas y modelos educativos, que cada vez más incluían tecnología en sus procesos.

Con base en lo anterior, se evidencia que en el campo de estudio de las evaluaciones estandarizadas de los aprendizajes existen áreas de oportunidad para continuar desarrollando conocimiento científico *in situ*; por citar un ejemplo, resulta evidente la ausencia de *softwares* educativos con capacidades de análisis psicométrico integradas, ya que en las investigaciones revisadas se encuentran estos aspectos separados; es decir, por un lado, se encuentra el *software* educativo que puede o no contar con capacidades evaluativas (esto dependiente del propósito del *software*, algunos son sólo de carácter informativo, comunicativo, etc.), y por otro, el *software* diseñado específicamente para realizar análisis psicométrico.

Sin embargo, se recalca de nueva cuenta que no existe la fusión de los mismos; como se abordó en la sección de medición de la calidad de la prueba, los cálculos que se efectúan son un tanto complejos y, probablemente, al considerar que se trata de pruebas estandarizadas (o de gran escala) se tenga la creencia de que incluir la capacidad psicométrica en un *software* que permita representar digitalmente una prueba lo haga muy pesado o pueda restarle rendimiento. Sin embargo, con los avances tecnológicos en cuanto a plataformas para desarrollo de *software*, gestores

de bases de datos y *hardware* de alto rendimiento, lo anterior no debería ser una limitante.

A la luz del resultado de esta revisión, se invita a los investigadores y especialistas en el campo a tomar o considerar como objeto de estudio el *software* educativo con capacidades de análisis psicométrico integradas. 

Referencias

- Aignerren, M. (1999). Análisis de contenido: una introducción. *La Sociología en sus escenarios*, (3). Recuperado a partir de <https://revistas.udea.edu.co/index.php/ceo/article/view/1550>
- Alcaraz, N. (2015). Evaluación versus calificación. *Aula de Encuentro*, 2(17), 209-236.
- Aliaga, J. (2006). Psicometría: test psicométricos, confiabilidad y validez. En A. Quintana, *Psicología: Tópicos de Actualidad* (pp. 85-108). Lima: Universidad Nacional Mayor de San Marcos.
- Almaguel, A., Álvarez, D. y Pernía, L. A. (2016). Software educativo para el trabajo con matrices. *Revista Digital: Matemática, Educación e Internet*, 16(2), 1-12.
- Argibay, J. (2006). Técnicas psicométricas. Cuestiones de validez y confiabilidad. *Subjetividad y procesos cognitivos*, 8, 15-33.
- Árraga, M. y Sánchez, M. (2012). Validez y confiabilidad de la Escala de Felicidad de Lima en adultos mayores venezolanos. *Universitas Psychologica*, 21(2), 381-393.
- Attorresi, H., Lozzia, G., Abal, F., Galibert, M. y Aguerri, M. (2009). Teoría de Respuesta al Ítem. Conceptos básicos y aplicaciones para la medición de constructos psicológicos. *Revista Argentina de Clínica Psicológica*, 18(2), 179-188.
- Backhoff, E. (2018). Evaluación estandarizada de logro educativo: contribuciones y retos. *Revista Digital Universitaria*, 19(6), 1-14.
- Backhoff, E., Larrazolo, N., y Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *REDIE. Revista Electrónica de Investigación Educativa*, 2(1), 11-28.
- Baladrón, J., Sánchez, F., Romeo, J., Curbelo, J., Villacampa, P., y Jiménez, P. (2018). Evolución de los parámetros dificultad y discriminación en el ejercicio de examen MIR. Análisis de las convocatorias de 2009 a 2017. *FEM: Revista de la Fundación Educación Médica*, 21(4), 181-193.

- Bogantes, J. (2015). Estrategias para la evaluación en educación a distancia: un análisis de las opciones empleadas en el programa de educación general básica de la UNED. *Innovaciones educativas*, 17(22), 15-25.
- Brooks, G. y Johanson, G. (2003). TAP: Test Analysis Program. *Applied Psychological Measurement*, 27(4), 303-304.
- Castells, M. (2010). La sociedad red: una visión global. *Enl@ce: revista venezolana de información, tecnología y conocimiento*, 7(1), 139-141.
- Cataldi, Z. (2000). *Una metodología para el diseño, desarrollo y evaluación de software educativo*. Tesis de Magister. Buenos Aires: Universidad Nacional de La Plata (Argentina).
- Cataldi, Z., Lage, F., Pessacq, R. y García, R. (1999). Ingeniería de software educativo. Ponencia presentada en el congreso “V Congreso Internacional de ingeniería informática”, Buenos Aires. Recuperado de <http://laboratorios.fi.uba.ar/lsi/c-icie99-ingenieriasoftwareeducativo.pdf>
- Cechova, I., Neubauer, J. y Sedlacik, M. (2014, octubre 30-31). *Computer-adaptive testing: item analysis and statistics for effective testing* [Ponencia]. European Conference on e-Learning, Copenhagen, Dinamarca.
- Contreras, G. (2010). Diagnóstico de dificultades de la evaluación del aprendizaje en la universidad: un caso particular en Chile. *Educación y Educadores*, 13(2), 219-238.
- Cortada, N. (2004). Teoría de respuesta al ítem: supuestos básicos. *Revista Evaluar*, 4(1), 95-110.
- Couturejuzón, L. (2003). Cumplimiento de los principios didácticos en la utilización de un software educativo para la educación superior. *Educación Médica Superior*, 17(1), 53-57.
- Cova, Á., Arrieta, X. y Riveros, V. (2008). Análisis y comparación de diversos modelos de evaluación de software educativo. *Enl@ce: Revista Venezolana de Información, Tecnología y Conocimiento*, 5(3), 45-67.
- Díaz, G., Pérez, M., Mendoza, L. y Grimán, A. (2003, noviembre 24-1 28). *Calidad Sistémica del Software Educativo* [Ponencia]. Congreso Internacional Edutec' 2003: Gestión de las Tecnologías de la Información y la Comunicación en los diferentes ámbitos educativos. Universidad Central de Venezuela, Caracas, Venezuela.
- Esterkin, V. y Pons, C. (2017). Evaluación de calidad en el desarrollo de software dirigido por modelos. *Ingeniare. Revista chilena de ingeniería*, 25(3), 449-463.
- Etchezahar, E., Prado-Gascó, V., Jaume, L. y Brussino, S. (2014). Validación argentina de la Escala de Orientación a la Dominancia Social. *Revista Latinoamericana de Psicología*, 46(1), 35-43.

- Ezpeleta, L., de la Osa, N., Domenech, J. M., Navarro, J. y Losilla, J. (1997). Fiabilidad test-retest de la adaptación española de la Diagnostic Interview for Children and Adolescents (DICA-R). *Psicothema*, 9(3), 529-539.
- Fernández, M., Alcaraz, N. y Sola, M. (2017). Evaluación y pruebas estandarizadas: Una reflexión sobre el sentido, utilidad y efectos de estas pruebas en el campo educativo. *Revista Iberoamericana de Evaluación Educativa*, 10(1), 51-67.
- Ferreira, M. y Backhoff, E. (2016). Validez del Generador Automático de Ítems del Examen de Competencias Básicas (Excoba). *RELIEVE-Revista Electrónica de Investigación y Evaluación Educativa*, 22(1), 1-16.
- Galvis, A. (2000). *Ingeniería de software educativo*. Colombia: Ediciones Uniandes.
- García, E., Vite, O., Navarrate, M. A., García, M. Á. y Torres, V. (2016). Metodología para el desarrollo de software multimedia educativo MEDESME. *CPU-e. Revista de Investigación Educativa*, 23, 216-226.
- Gómez, M. T. (1997 octubre 27-29). *Un ejemplo de evaluación de software educativo multimedia* [Ponencia]. Congreso Internacional EDUTEC 97: Creación de materiales para la innovación educativa con nuevas tecnologías. Málaga, España.
- Gómez, R. L. (2004). Calidad educativa: más que resultados en pruebas estandarizadas. *Revista educación y pedagogía*, 16(38), 75-89.
- González, Y. L. (2006). *Diseño e implementación de un software educativo para el autoaprendizaje del diseño de bases de datos relacionales* [Tesis de licenciatura]. Universidad Nacional Autónoma de México.
- Hidalgo, N. y Murillo, F. J. (2017). Las concepciones sobre el proceso de evaluación del aprendizaje de los estudiantes. *REICE: Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, 15(1), 107-128.
- Izquierdo, B. (2008). De la evaluación clásica a la evaluación pluralista. Criterios para clasificar los distintos tipos de evaluación. *EMPIRIA. Revista de Metodología de las Ciencias Sociales*, 16, 115-134.
- Jiménez, A. (2004). *El estado del arte en la investigación en Ciencias Sociales*. Red de Bibliotecas Virtuales de CLACSO.
- Jornet, J. M. (2017). Evaluación estandarizada. *Revista iberoamericana de evaluación educativa*, 10(1) 1, 5-8.
- López, F. (2002). El análisis de contenido como método de investigación. *Revista de Educación*, 4, 167-179.
- López, J., Hernández, C. y Farran, Y. (2011). Una plataforma de evaluación automática con una metodología efectiva para la enseñanza/aprendizaje en

- programación de computadores. *Ingeniare. Revista Chilena de Ingeniería*, 19(2), 265-277.
- Madariaga, C. J., Rivero, Y. y Leya, A. R. (2016). Propuesta metodológica para desarrollo de software educativo en la Universidad de Holguín. *Ciencias Holguín*, 22(4), 1-17.
- Mandeville, P. (2005). El coeficiente de correlación intraclase (ICC). *Ciencia UANL*, 8(3), 414-416.
- Marquès, P. (1998). La evaluación de programas didácticos. *Comunicación y Pedagogía*, 149, 53-58.
- Marr, L., Gupchup, G. y Anderson, J. (2012). An evaluation of the psychometric properties of the Purdue Pharmacist Directive Guidance Scale using SPSS and R software packages. *Research in Social and Administrative Pharmacy*, 8(2), 166-171.
- Martínez, F. (2001). Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate. *Revista de la educación superior*, 30(120), 1-12.
- Martínez, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista electrónica de investigación educativa*, 11(2), 1-18.
- Martínez, M. y Sauleda, N. (1992). La evaluación de software educativo en el escenario de la evolución de los paradigmas educativos. *Enseñanza y Teaching: Revista Interuniversitaria de Didáctica*, 10, 161-174.
- Medina, J., Ramírez, M. H. y Miranda, I. (2019). Validez y confiabilidad de un test en línea sobre los fenómenos de reflexión y refracción del sonido. *Apertura: Revista de Innovación Educativa*, 11(2), 104-121.
- Mendivil, T. N. (2012). Sistema de evaluación del aprendizaje en los estudiantes de educación superior en la región caribe colombiana. *Dimensión empresarial*, 10(1), 100-107.
- Mora, A. (2004). La evaluación educativa: Concepto, períodos y modelos. *Actualidades investigativas en educación*, 4(2), 1-28.
- Moreno, T. (2016). Las pruebas estandarizadas en la escuela contemporánea, ¿llave o cerrojo para la mejora de la educación? *Temas de Educación*, 22(1), 83-96.
- Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo: Revista del Colegio Oficial de Psicólogos*, 31(1), 57-66.
- Navarro, E. (1999). Análisis de productos multimedia educativos. *Comunicación y Pedagogía*, 157, 37-40.

- Orjuela, A. y Rojas, M. (2008). Las metodologías de desarrollo ágil como una oportunidad para la ingeniería del software educativo. *Revista Avances en Sistemas e Informática*, 5(2), 59-171.
- Pérez, J. H., Acuña, N. y Arratia, E. R. (2008). Nivel de dificultad y poder de discriminación del tercer y quinto examen parcial de la cátedra de citohistología 2007 de la carrera de medicina de la UMSA. *Cuadernos Hospital de Clínicas*, 53(2), 16-22.
- Poole, B. (1999). *Tecnología Educativa: educar para la sociedad de la comunicación y del conocimiento* (2da. ed.). España: Mc Graw Hill.
- Popham, J. (1999). Why standardized test don't measure educational quality (Programa de Promoción de la Reforma Educativa en América Latina y el Caribe, Grupo de Trabajo sobre Estándares y Evaluación, Trans). *Educational Leadership*, 56(6), 2-11.
- Quero, M. (2010). Confiabilidad y coeficiente Alpha de Cronbach. *Telos*, 12(2), 248-252.
- Ravela, P. (2010). ¿Qué pueden aportar las evaluaciones estandarizadas a la evaluación en el aula? Programa de Promoción de la Reforma Educativa en América Latina y el Caribe. Preal. *Serie Documentos*, 47, 3-25.
- Reidl, L. M. (2013). Confiabilidad en la medición. *Investigación en educación médica*, 2(6), 107-111.
- Reyes, F., Fernández, F. y Duarte, J. (2015). Herramienta para la selección de software educativo aplicable al área de tecnología en educación básica. *Entramado*, 11(1), 186-193.
- Santelices, M. y Valenzuela, F. (2015). Importancia de las características del profesor y de la escuela en la calidad docente: Una aproximación desde la Teoría de Respuesta del Ítem. *Estudios pedagógicos*, 41(2), 233-254.
- Serra, A. y Peña, J. (2006). Fiabilidad test-retest e interevaluador del Test Barcelona. *Neurología*, 21(6), 277-281.
- Shepard, P. (2006). La evaluación en el Aula. En R. Brennan (Ed.), *Educational Measurement* (pp. 623-646). Praeger Westport.
- Squires, D. y McDougall, A. (1997). *Cómo elegir y utilizar software educativo: guía para el profesorado*. Madrid: Ediciones Morata.
- Straccia, L., Zanetti, P. y Pollo, M. F. (2019, octubre 14-18). Definición de un estándar para la evaluación de calidad de software educativo [Ponencia]. XXV Congreso Argentino de Ciencias de la Computación. Córdoba, Argentina.
- Thoe, N., Fook, F. S. y Thah, S. S. (2009). Use of ICT tool for Item Analysis of a Science Performance Test. *Journal of Educational Technology*, 9(1), 5-15.

- Tirado, F., Backhoff, E. y Larrazolo, N. (2016). La revolución digital y la evaluación: un nuevo paradigma. *Perfiles educativos*, 38(152), 182-201.
- Tiramonti, G. (2014). Las pruebas PISA en América Latina: resultados en contexto. *Avances en Supervisión Educativa*, 20, 1-24.
- Tristán, A. y Pedraza, N. (2017). La objetividad en las pruebas estandarizadas. *Revista Iberoamericana de evaluación educativa*, 10(1), 11-31.
- Ulloa, C. (2015). *Análisis de contenido*. Recuperado de <https://bit.ly/3bULQw7>
- Umaña, A. C., Calvo, X. y Salas, N. (2017). Evaluar para aprender: estado actual de catorce asignaturas en la universidad estatal a distancia de Costa Rica. *Revista Electrónica Calidad en la Educación Superior*, 8(2), 24-61.
- Urrútia, G. y Bonfill, X. (2010). Declaración PRISMA: Una propuesta para mejorar la publicación de revisiones sistemática y metaanálisis. *Medicina Clínica*, 135(11), 507-511.
- Vidal, M., Gómez, F. y Ruiz, A. (2010). Software educativos. *Educación Médica Superior*, 24(1), 97-110.
- Zúñiga, M., Solar, M. I., Lagos, J., Báez, M. y Herrera, R. (2014). Evaluación de los aprendizajes: un acercamiento en educación superior. En CINDA-Centro Interuniversitario de Desarrollo, *Evaluación del aprendizaje en innovaciones curriculares de la educación superior* (pp. 15-38). Santiago de Chile: Ediciones e Impresiones Copygraph.