DOI: https://doi.org/10.56712/latam.v4i5.1397

El concepto de validez y el enfoque basado en argumentos para un examen de ingreso a la universidad

The concept of validity and the argument-based approach for a university entrance assessment

Karla Karina Ruiz Mendoza
ruiz.karla32@uabc.edu.mx
https://orcid.org/0000-0001-8978-8364
IIDE de la Universidad Autónoma de Baja California
Ensenada – México

Luis Horacio Pedroza Zúñiga
horacio.pedroza@uabc.edu.mx
https://orcid.org/0000-0002-5256-2967
IIDE de la Universidad Autónoma de Baja California
Ensenada – México

Artículo recibido: 11 de noviembre de 2023. Aceptado para publicación: 27 de noviembre de 2023.
Conflictos de Interés: Ninguno que declarar.

Resumen

La Universidad Autónoma de Baja California (UABC) está en un proceso crítico de desarrollo de su examen de ingreso, lo que requiere un escrutinio riguroso respecto a la validez y validación de dicho examen. La validez, un concepto con una rica historia y evolución, ha transitado por diversas fases conceptualizadoras hasta llegar a un enfoque basado en argumentos. A través de los tiempos, desde el auge de los test de inteligencia y pruebas psicológicas, hasta las elaboraciones de estándares de pruebas en la década de los cincuenta, el concepto de validez ha ido refinándose, pasando por la unificación del concepto por Messick en los años 70-90, hasta llegar a la deconstrucción de la validez en el periodo 2000-2012. Michael Kane, con su enfoque basado en argumentos, ha modificado la perspectiva de estudio de la validez, enfocándose en el "cómo" más que en el "qué", proponiendo dos tipos de argumentos: el Argumento de Interpretación o Uso (IUA) y el argumento de validez. Este último incluye interpretaciones y usos de los puntajes del test, apoyado por cuatro tipos de inferencias: puntuación, generalización, extrapolación e implicaciones. La estructura del argumento de validez, apreciable en pruebas como el TOEFL, se sugiere como una guía viable para la validación de exámenes de ingreso a la universidad, ajustando sus inferencias y garantías a las especificidades del área de conocimiento y habilidades a ser evaluadas. En este sentido se busca destacar la imperante necesidad de un enfoque sistemático y bien fundamentado en la construcción de pruebas y exámenes, especialmente para instituciones educativas que buscan garantizar una evaluación precisa y válida de sus futuros estudiantes.

Palabras clave: validez, validación, validación por argumento, TOEFL, educación superior

Abstract

The Autonomous University of Baja California (UABC) is in a critical process of developing its entrance exam, which requires rigorous scrutiny regarding the validity and validation of said exam. Validity, a concept with a rich history and evolution, has gone through various conceptual phases until reaching an argument-based approach. Over time, from the rise of intelligence tests and psychological assessments to the elaboration of testing standards in the fifties, the concept of validity has been refined, passing through the unification of the concept by Messick in the 70s-90s, to the deconstruction of validity in the period 2000-2012. Michael Kane, with his argument-based approach, has modified the perspective of validity study, focusing on the "how" rather than the "what", proposing two types of arguments: the Interpretation or Use Argument (IUA) and the validity argument. The latter includes interpretations and uses of the test scores, supported by four types of inferences: scoring, generalization, extrapolation, and implications. The structure of the validity argument, noticeable in tests like the TOEFL, is suggested as a viable guide for the validation of university entrance exams, adjusting its inferences and warranties to the specificities of the area of knowledge and skills to be evaluated. In this sense, there is a push to highlight the pressing need for a systematic and well-founded approach in the construction of tests and examinations, especially for educational institutions that seek to guarantee an accurate and valid evaluation of their future students.

Keywords: validity, argument validity, TOEFL, university admissions, standards

Como citar: Ruiz Mendoza, K. K., & Pedroza Zúñiga, L. H. (2023). El concepto de validez y el enfoque basado en argumentos para un examen de ingreso a la universidad. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades 4(5), 1346 – 1359. https://doi.org/ 10.56712/latam.v4i5.1397*

INTRODUCCIÓN

En la actualidad la Universidad Autónoma de Baja California (UABC) se encuentra en un proceso para la elaboración de su examen de ingreso a la universidad, lo que requiere fina atención en el ámbito de validez y validación. Debido a todo lo anterior, el propósito de este trabajo es realizar una revisión sobre la definición del concepto de validez a través del tiempo, para entender cómo hemos llegado hasta el enfoque basado en argumentos y cómo lograr modelar cuando se desea elaborar una prueba o examen de ingreso a la universidad o en la comprobación de ciertos conocimientos idiomáticos. Por lo que, este texto se compone de cinco apartados: línea temporal del concepto validez; validez y estándares; una aproximación desde Kane al enfoque basado en argumentos para la validación de una prueba; y un acercamiento a las pruebas TOEFL para proponer un modelo guía para la validación de exámenes de ingreso a la universidad.

Aunque existen diversas definiciones sobre la validez la mayoría concuerda en que los elementos de una evaluación son representados adecuadamente en el instrumento del constructo a investigar (Delgado-Rico, Carrctero-Dios & Willibald, 2012). Kerlinger y Lee (2001) resumen muy bien la concepción básica de la validez en una prueba: “¿estamos midiendo lo que creemos que estamos midiendo?” (p.), es decir, la validez ayuda a determinar la relevancia de la evidencia para respaldar un instrumento e incluso exámenes o bien si el instrumento cumple el objetivo para el cual fue elaborado (Urrutia, et al., 2014). Por lo tanto, el concepto de fiabilidad o confiabilidad acompañará las formas de determinar la calidad de un instrumento (Urrutia, et al., 2014; Kerlinger y Lee, 2001). Así, la validez es parte de los procesos de estandarización de exámenes o instrumentos de medición, lo cual se ha vuelto una práctica frecuente para mejorar y tratar de predecir la toma de decisiones en educación (López & Douglas Willms, 2019; AERA, APA y NCME, 2014).

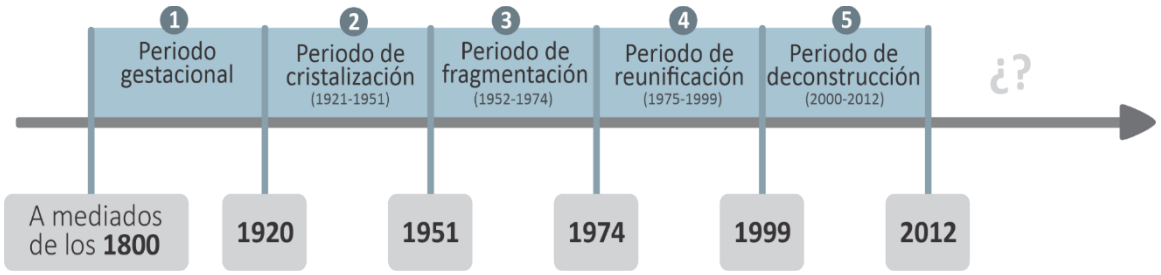
Por otra parte, la validez puede ser aplicada desde la investigación cualitativa o cuantitativa, en el primer tipo de investigación podríamos basarnos en la honestidad o en el grado de interés del investigador, mientras que en el segundo hablamos del uso de instrumentos adecuados e incluso del análisis estadístico de los datos (Cohen, 2007). En este sentido, la validez es parte fundamental de una investigación efectiva, y además fundamental cuando se trata del estudio de comportamientos (Cohen, 2007; Kerlinger y Lee, 2001). A esto podríamos agregar que algunos autores clasifican algunos tipos de validez (aunque esto no es correcto, se explica más adelante): de contenido, de criterio, de constructo, interna, externa, concurrente, aparente, de jurado, predictiva, consecuencial, sistémica, catalítica, ecológica, descriptiva, interpretativa, teórica y evaluativa (Cohen, 2007). O bien, de contenido, de criterio, de constructo, o por método multirrasgo-multimétodo (Kerlinger y Lee, 2001).

LÍNEA TEMPORAL DEL CONCEPTO DE VALIDEZ

La validez mantiene su propia historia, para explicar un poco sobre la evolución del concepto de validez retomaremos a Newton y Shaw (2014) quienes se preguntan qué es reivindicar la validez y cómo puede fundamentarse esta reivindicación; para ello elaboran un análisis desde la teoría de la validez, e identifica a Messick (1989), desde 1970 hasta los años de 1990, como un teórico que profundizó extensivamente en el tema pero que dejó poca claridad al respecto; el cual es retomado por García Median et al. (2017) en un capítulo sobre la evolución del concepto de validez. Entonces, como se observa en la figura 1, el concepto de validez ha pasado por periodos importantes dentro de su conceptualización, desde el cómo se va gestando hasta el periodo deconstructivo donde nos volvemos a realizar las preguntas sobre si lo que pensamos que es lo es.

Figura 1

Línea del tiempo del concepto validez



Fuente: Elaboración propia a partir de la traducción del capítulo An outline of the history of validity en Validity in Educational and Psychological Assessment de Newton y Shaw (2014).

Sobre el periodo gestacional (A gestational period, mid-1800s-1920) mencionan, Newton y Shaw (2014), que hubo un avance en la metodología de la estadística que aceleraron la estructuración de los exámenes basados en la premisa en probar los alcances de la mente humana. Ciertamente hubo un gran interés por el desarrollo de pruebas o test de inteligencia, también denominado test para la medición del coeficiente intelectual (CI) (Watson, 2002). Lo que trajo consigo el boom de la medición (de inteligencia, aptitudes, entre otros), sobre todo en Estados Unidos de América y Europa, en el caso de los norteamericanos utilizaron ciertas pruebas en la selección de los reclutados durante la Primera Guerra Mundial (Newton y Shaw, 2014; Watson, 2002).

Del periodo de cristalización (A period of crystallization, 1921-1951) se habla de un uso desmesurado de las pruebas para juzgar a los estudiantes pudiendo diagnosticar su "retraso" (backwardness) o bien su "excelencia" (excellence), debido a esto, naturalmente algunos miembros de la comunidad científica educativa comenzaron a cuestionar sus usos e intenciones (Newton & Shaw, 2014). Así, en 1921 la North American National Association of Directors of Educational Research publicaron sus intenciones para llegar a un consenso sobre el llamado "movimiento de medición" (measurement movement). Para este momento, entonces, surgieron dos aproximaciones: el análisis lógico del contenido; y, basado en la evidencia empírica de la correlación, es decir sobre la prueba y lo que se supone debe medir.

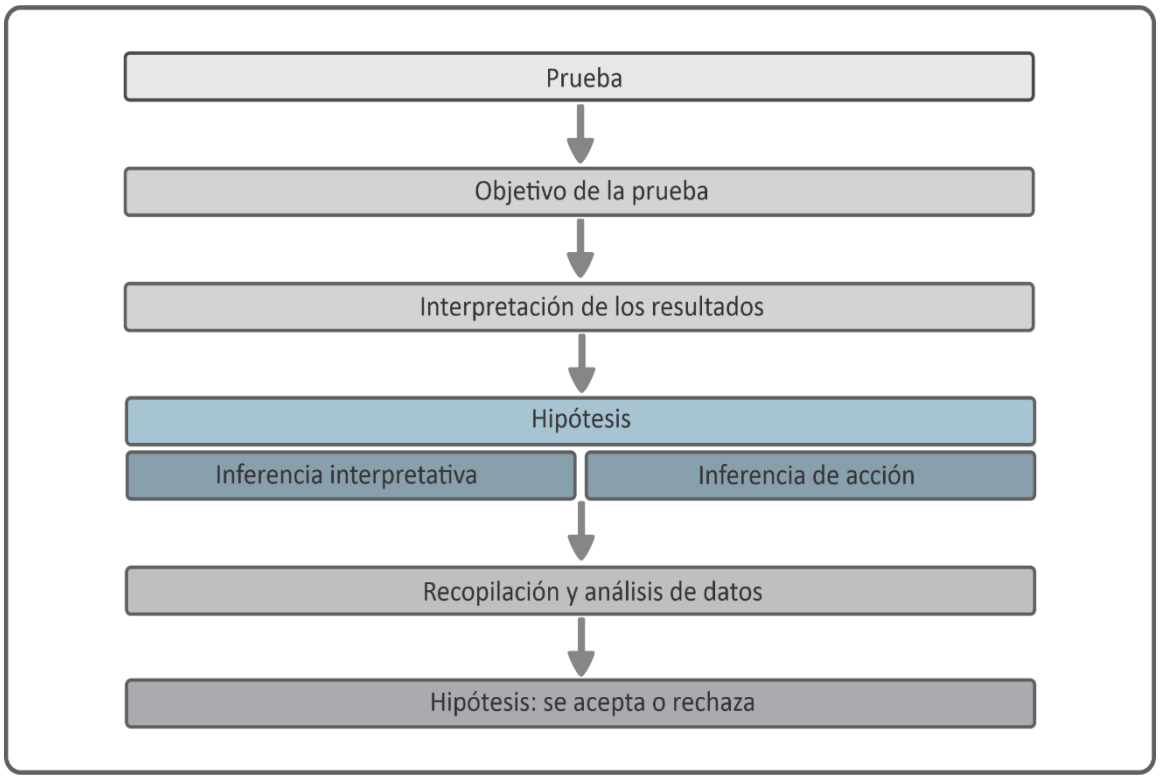
Hacia los cincuentas (con la participación concurrente de Lee Cronbach, 1951), debido al interés por estandarizar y establecer ciertos parámetros, la primera publicación en vías de estandarización de pruebas fue en 1954, llamado Recomendaciones Técnicas para las Pruebas Psicológicas y las Técnicas de Diagnóstico, elaborados y publicados por la APA, aunque hubo un borrador hacia 1952 (Newton y Shawn, 2014). Después la AERA y la National Council on Measurement Used in Education (NCMUE) elaboran las Recomendaciones Técnicas para Pruebas de Rendimiento y publicadas por la National Education Association (NEA) hacia 1955 (APA, AERA, NCME, 2014). Para 1966 la AERA, la APA y el NCME sustituyeron estas para dar paso a la primera edición de los Estándares para Pruebas Educativas y Psicológicas; existiendo publicaciones hacia 1974, 1985, 1999 y la actualización en español del año 2014 (APA, AERA, NCME, 2014). En este sentido, para este mismo año (1966) debido a la revisión realizada, se puede observar la fragmentación del concepto validez, es decir, se comienzan a clasificar en tres tipos de validez: de contenido, relacionados con los criterios y de constructo.

El periodo de re-unificación de la validez (The [re]unification of validity [1974-1999]) se le conoce como los "años de Messick" (Messick Years) debido a que logró unificar el concepto de validez poniendo la medición como el centro de la validez en cualquier contexto, en este sentido, la validez depende de la suma de evidencias posibles sobre la prueba y no de un solo tipo de evidencia (Newton y Shawn,

20149). En la visión de Messick la validez de constructo es el único existente, pues el objetivo es medir un constructo el cual depende del uso que se pretenda dar de la prueba, obsérvese la figura 2.

Figura 2

Marco de referencia de Messick



Fuente: Retomado, con cambios mínimos, de El concepto moderno de validez y su uso en educación médica, figura 1 de Carrillo, Sánchez y Leenen (2020).

Sobre el periodo de la deconstrucción de la validez (The deconstruction of validity, 2000-2012), hubo una gran influencia por parte de Messick en esta área teórica, así como la presencia de Michael Kane con su enfoque basado en argumentos, puesto que ahora se debe pensar en el cómo y no en el qué, es decir, cómo procedemos a validar en fragmentos pequeños y entendibles, a diferencia de las antiguas discusiones sobre entender qué era la validez en sí misma. Chapelle (2021), por ejemplo, diría que Kane define a la validez a partir de entender la validación, por ello es importante, en una nueva etapa, partir del enfoque basado en argumentos y, a la vez, estructurar la validación de una prueba. A continuación, seguiremos analizando la validez y los estándares oficiales para definir completamente el qué y posteriormente cuestionar al cómo con Kane.

VALIDEZ Y LOS ESTÁNDARES

En lo que respecta a los Estándares para Pruebas Educativas y Psicológicas (2014; de ahora en adelante le llamaremos Estándares) de la American Educational Research Association (AERA), American Psychological Association (APA), y la National Council on Measurement in Education (NCME), se define a la validez como "(...) grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para usos propuestos de las pruebas." (p.11), en este sentido, cuando hablamos de validez hablamos de un concepto unitario, es decir, una acumulación de evidencias que respalda la interpretación en los puntajes propuestos; es por ello que es erróneo definir

una tipología de validez (APA, AERA y NCME, 2014). Por ende, hablamos de tipos de evidencia de validez, las cuales son las siguientes (APA, AERA y NCME, 2014):

Evidencia basada en el contenido de la prueba: el punto relevante en este tipo de evidencia es el contenido, el cual "hace referencia a los temas, la redacción y el formato de los ítems" (pp.14-15). En este caso hay que tener sumo cuidado en la elaboración de los ítems así como en la clasificación, por lo que aquí se sugiere el juicio de expertos. En educación a este tipo de evidencia se le conoce como "(...) alineación, que involucra evaluar la correspondencia entre estándares de aprendizaje para estudiantes y el contenido de la prueba." (p.15).

Evidencia basada en los procesos de respuesta: el enfoque es al proceso de respuestas de los participantes, incluye preguntarles a los examinandos sobre sus estrategias para la resolución del problema, también es posible monitorear el tiempo de respuesta o el movimiento de los ojos, este tipo de evidencias pueden ayudar a comprender "(...) en qué medida las capacidades irrelevantes o auxiliares al constructo pueden influir de manera diferencial en el desempeño de los examinandos en la prueba." (p.16).

Evidencia basada en la estructura interna: indican el grado de relación entre los ítems y los componentes de la prueba para saber si se alinean al constructo y por lo tanto a las interpretaciones de los puntos dados en la prueba.

Evidencia basada en relaciones con otras variables: este tipo de evidencia sugiere la relación entre variables con otras variables externas para lograr un análisis de los puntajes; es decir, es una evidencia que busca coherencia. Sobre este tipo de evidencias podemos sub tipificarlos en tres: evidencia convergente (evidencia de un mismo constructo o similar para determinar puntajes) y discriminante (se da en la relación de los puntos de la prueba y medidas de constructos diferentes); Relaciones prueba-criterio (dependerá de la confiabilidad, relevancia y validez de la interpretación); y, Generalización de validez (comúnmente refiere a los metaanálisis o estudios estadísticos que ayuden a generalizar un criterio).

Evidencia de validación y consecuencias de las pruebas: se analizan las consecuencias probables de las pruebas, por lo que se debe realizar un análisis de las consecuencias de las consecuencias, si bien se pueden adquirir beneficios y una brújula para la toma de decisiones en instituciones o escuelas, hay que ser cautelosos (indica la AERA, APA & NCME, 2014). Así, toman a discusión: la interpretación y usos de puntajes de la prueba previstos por los desarrolladores de la prueba; Afirmaciones hechas sobre el uso de la prueba que no se basan directamente en interpretaciones de los puntajes de la prueba; y, consecuencias que son imprevistas.

En la actualidad no se ha publicado una nueva versión de los estándares, es decir, sigue siendo la edición 2014 la que describimos en este texto, además, su última versión ha sido la traducción al español en 2018. La teoría de la validez de pruebas también sigue manteniendo la versión de Samuel Messick (1989) y de Michael Kane (2006) en esta versión de la AERA, APA y NCME (2014), no obstante, al día de hoy se prefiere la versión del enfoque basado en argumentos de Kane por su versatilidad (Carrillo, Sánchez y Leenen, 2020; Pedrosa, Suárez-Álvarez y García-Cueto, 2014; Cook, et al., 2015).

EL ENFOQUE BASADO EN ARGUMENTOS DE KANE

Kane se encuentra en el periodo de deconstrucción (2000-2012), también nombrado como validación moderna, en este sentido, Kane modifica la forma en la que se ha enfocado el estudio de la validez del qué de Messick al cómo. Para Kane (2013) no solo se puede interpretar la puntuación desde un

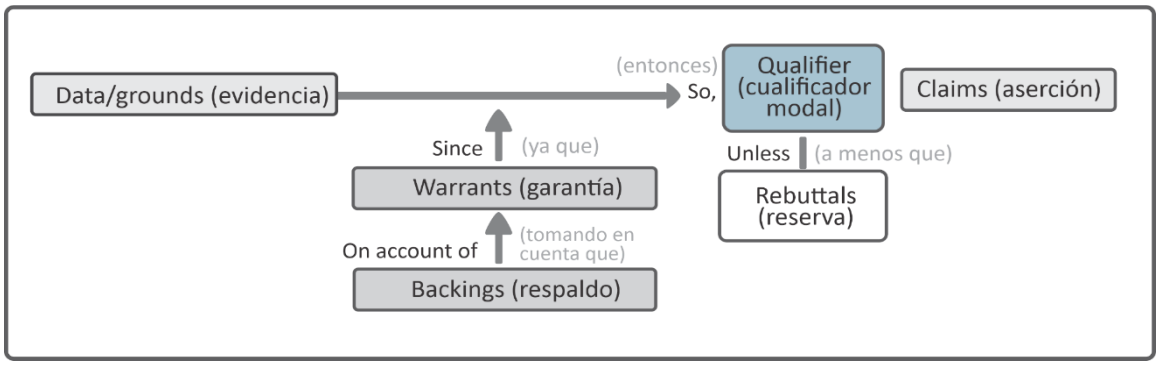
constructo[1], si no de la interpretación de un argumento (Chapelle et al., 2010; Chapelle, 2021). El enfoque por argumentos es simple: se expone lo afirmado y se evalúa eso que se afirma, "The approach is quite simple: state what is being claimed and evaluate the claims being made." (Kane, 2013, p.451).

Así, Kane (2006, 2011, 2013) propuso dos tipos de argumentos, el primero basado en la lógica del modelo de Toulmin (1958), al cual denominó como el Argumento de Interpretación o Uso (IUA, por sus siglas en inglés); el IUA especifica las aserciones a evaluar, es decir, las suposiciones relacionadas a las interpretaciones y usos de las puntuaciones de las pruebas (o test). El segundo es sobre el argumento de validez (validity argument) el cual propone la interpretación y el uso que se dispone a los puntajes del test, al que acompañan cuatro tipos de inferencias que se explicarán más adelante (Taylor, 2013); que en cierta medida también se encuentra acompañado del modelo de Toulmin (1958). Estas interpretaciones resultantes se pueden tomar desde dos perspectivas, ya mencionadas, uno por el significado de las puntuaciones de la prueba y dos, lo que se hace con los resultados de la prueba (Taylor, 2013; Carrillo, Sánchez y Leenen, 2020).

Asimismo, cabe recordar que Kane colaboró en los Estándares de la edición 2014 (AERA, APA & NCME), donde definen al argumento de validez como: "Justificación explícita del grado en que la evidencia acumulada y la teoría respaldan una interpretación propuesta de los puntajes para el uso previsto." (p. 242), por lo que las inferencias tienden a necesitar de diversas herramientas y evidencias pero que no necesariamente se deben cumplir todas (Kane, 2013; Carrillo, Sánchez y Leenen, 2020).

Figura 3

Modelo de Toulmin



Fuente: Retomado de Toulmin (1958) con las traducciones al español por Rodríguez (2004).

Para entender un poco más sobre la validez del argumento, hay que comprender el modelo de Toulmin (1958). Éste parte de las argumentaciones cotidianas, las cuales son inferencias en su forma "si-entonces" y que en su forma básica se compone de una aserción, datos y una garantía; por ejemplo: si presiono el botón de encender entonces se iluminará el cuarto (aserción); cada vez que se presiona el botón encender se enciende una luz. En la figura 3 se puede visualizar el modelo para comprender mejor este tipo de argumentación, no obstante, debido a que existen algunas discrepancias en la traducción del modelo de Toulmin (Rodríguez, 2004) se utilizarán los términos en su idioma original (inglés) y con la traducción entre paréntesis (también véase la figura 3).

[1] Recordando que un constructo se refiere al concepto o característica y que para medirlos diseñamos una prueba (AERA, APA y NCME, 2014).

Data/grounds (Evidencia): es una afirmación o inferencia previa (Toulmin, 1958; Toulmin, Rieke y Janik, 1984).

Warrant (garantía): regla que nos ayuda a inferir una afirmación (Kane, 2013).

Backing (respaldo): es el soporte de la garantía (warrant), una justificación o bien una razón para regresar a la garantía (Kane, 2013, 2011; Taylor, 2014).

Qualifier (cualificador modal): es la probabilidad de la aserción (Rodríguez, 2004).

Claim (aserción o afirmación): tesis a defender o asunto a debatir (Rodríguez, 2004).

Rebuttal (reserva): son las objeciones que posiblemente se puedan formular (Kane, 2013).

Argument (extra): es una serie de inferencias vinculadas (Kane, 2013).

En el caso del ejemplo de encender la luz a partir de un botón, podríamos decir que la garantía es que es un común denominador en los encendedores de luz; el respaldo de la garantía sería que si lo dice el instructivo entonces así debe suceder; el dato es que siempre que otra persona presiona ese botón se enciende la luz o bien que así está constituido el dispositivo; el cualificador modal es que lo más probable es que encienda; la reserva es que a menos que no funcione el aparato o no haya luz en la colonia o algo parecido entonces no encendería.

Por otra parte, Kane (2011) define tres criterios básicos para la interpretación de argumentos que son importantes al trabajar con inferencias: claridad del argumento (clarity of the argument), aquí se deben detallar la garantía y el respaldo; coherencia del argumento (coherence of the argument), se propone que el razonamiento lleve con facilidad a persuadir al otro; y, la plausibilidad (o verosimilitud) de las inferencias y supuestos (plausibility of inferences and assumptions), puede basarse en hipótesis ya dadas por sentadas, por documentación y análisis profundo o bien, por pruebas empíricas.

Tabla 1

Las inferencias de Kane

Inferencia	Refiere a	Garantía o procedimientos a seleccionar
Puntuación (scoring)	Se toman información de los datos de los puntajes como afirmación, se establecen criterios y reglas.	Reglas o rúbricas de puntuación. Estandarización de puntajes. Juicio de grupos de expertos.
Generalización (generalization)	La puntuación observada del evaluado nos da una estimación de la puntuación del universo evaluado.	Teoría de la generalizabilidad (teoría G). Tamaño de la muestra y cantidad de preguntas.
Extrapolación (extrapolation)	Estimación de la función de la prueba en el contexto real.	Análisis entre la relación de la prueba y su función en otros contextos. Ecuación de regresión.
Implicaciones (Implications)	Interpretación y toma de decisiones	Aprobación o no aprobación del estándar. Acciones. Consecuencias voluntarias o involuntarias.

Fuente: Elaboración propia a partir de Kane (2001, 2013); Cook et al, 2015; Carrillo, Sánchez y Leenen, 2020.

Entonces, Kane utiliza el modelo de Toulmin para reforzar la aserción y sus argumentaciones a través de la coherencia y la interpretación:

The interpretive argument specifies the inferences involved in getting from the observed performances to the conclusions to be drawn and the decisions to be made based on test scores. It would include a network or chain of inferences. The validity argument would provide a critical appraisal of the coherence of the interpretive argument and of the warrants and backing for the inferences in this argument. (Kane, 2011, p.12)

En este sentido, el primer paso en una prueba es proponer el argumento de uso o interpretación (AUI) o bien redes de inferencias y suposiciones que, en un segundo momento, pueden ser apoyadas por las cuatro inferencias de Kane (2013b; Cook, et al., 2015): puntuación (scoring), generalización (generalization), extrapolación (extrapolation), e implicaciones (implications), en la tabla 1 se especifican cada uno de ellos.

Estas cuatro inferencias recopilan diferentes evidencias que se aportan a la validez. Cook, et al. (2015) lo describen como: primero hay que realizar la observación singular del puntaje, ya sea que éste sea por pregunta de opción múltiple o portafolio de evidencias u otro tipo de observación, después se usan estos puntajes para generar un puntaje general (generalización), seguido de ello se procede a la extrapolación, es decir, trazar las inferencias a partir de los puntajes obtenidos y lo que podrían implicar en la vida real, por último, analizar las implicaciones a través interpretar toda esta información para la toma de decisiones.

APROXIMACIONES A UN MODELO DE VALIDEZ PARA UN EXAMEN DE INGRESO A LA UNIVERSIDAD

Si bien en el ámbito médico psicología existe mayor aproximación sobre el tema de validez de pruebas o test (Pedrosa, et al., 2014; Cook, et al., 2015), Chapelle, et al., han profundizado en el Test of English as a Foreign Language, mejor conocido como TOEFL, en cuanto respecta al tema de validez. En 2008 publicaron el libro Building a Validity Argument for the Test of English as a Foreign Language y, en 2010, el artículo Does an Argument-Based Approach to Validity Make a Difference? En un primer momento hacen referencia a Messick y a los Estándares de la AERA, APA y NCME, pero sus interpretaciones y definiciones no fueron lo ideal para la interpretación de este tipo de prueba tan rigurosa, sobre todo porque hablan del manejo del idioma para entrar a universidades de habla inglesa por lo que revisan autores como Kane, Crooks y Cohen. Por otra parte, para lo que compete a este apartado, estos estudios nos pueden ayudar a definir si son o no aplicables a un examen de ingreso a la universidad; sin definir país o ciudad.

Así, Chapelle, et al., (2010) afirman que el modelo de Kane como se encontraba no era suficiente, es decir, con las inferencias mencionadas con anterioridad (puntuación, generalización, extrapolación, implicación), para lo que proponen cuatro inferencias más que ayuden a estructurar la prueba y quitar la inferencia de puntuación y de implicación, los agregados son: descripción del dominio, evaluación, explicación y utilización. Algunas de estas definiciones como la descripción del dominio y la generalización se encuentran disponibles en los Estándares (AERA, APA & NCME, 2014). Bajo estas seis inferencias Chapelle, et al. (2010), describen sus garantías y supuestos, aquí solo se presentan las inferencias y su interpretación:

Descripción del dominio: describir cuidadosamente y desarrollar los ítems para que reflejen el dominio (en cuanto a contenido, nivel cognitivo, o poder revisar de forma sistemática los errores) (Kane, 2004, tomado de Chapelle, et al., 2010).

Evaluación: si se coloca una puntuación de 1 el cual conlleva a un argumento interpretativo del qué significa ese punto, entonces debe evidenciarse a través de alguna rúbrica que demuestre los criterios de puntuación y la coherencia con la que se aplicó.

Generalización: es un término vinculado a la confiabilidad, ya que se basa en la selección de la puntuación en tareas similares.

Explicación: el comportamiento de medición debe ser congruente con la prueba; en el caso de Chapelle, et al., se explica que hubo una correlación en la medición entre el TOEFL y otras pruebas.

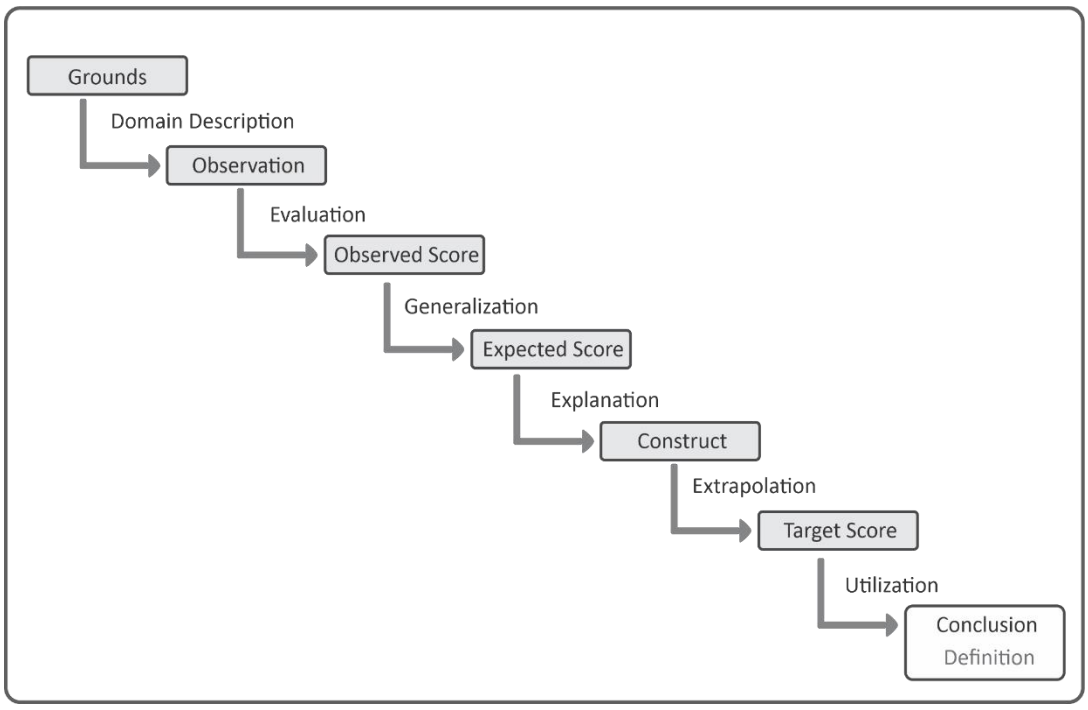
Extrapolación: refiere a si la prueba puede determinar el comportamiento después de la misma, es decir, en el caso del idioma inglés refiere a si el estudiante se desenvolverá de la misma manera en su contexto universitario.

Utilización: refiere a la fácil interpretación de la puntuación de los resultados del alumnado para la toma de decisiones (Bachman, 2005, tomado de Chapelle, et al., 2010).

De esta forma, Chapelle, et al. (2010), logran definir una estructura para el argumento de validez del TOEFL, obsérvese figura 4, donde la conclusión revela la habilidad aprendida (en este caso es el idioma inglés) y para concluir que esta habilidad se encuentra cubierta, el que presente la prueba deberá dominar habilidades de habla, escritura y escucha para el desenvolvimiento universitario. Además, los puntos deben ser útiles para la toma de decisión sobre su admisión.

Figura 4

Estructura del argumento de validez del TOEFL



Fuente: Elaboración basada en la figura 3 en Does an Argument-Based Approach to validity Make a Difference? de Chapelle, et al. (2010).

En este sentido, referimos a: grounds, como el objetivo es el dominio del uso de la lengua; observation: observación del conocimiento, capacidades y habilidades en situaciones representativas del dominio del lenguaje; observed score: observar si el puntaje es el esperado y/o son relevantes; expected score: puntuación esperada; construct: constructo a estudiar, el concepto a medir; target score: revisar si el puntaje objetivo se vincula con la conclusión y funciona para la toma de decisiones.

Finalmente, la figura 4 demuestra una síntesis de la estructura del argumento de validez del TOEFL, el cual podría ser una guía para la validación de exámenes de otros idiomas, sobre todo tratando de aprovechar la ya mencionada flexibilidad del modelo de Kane (2001, 2006, 2013). Por lo que se propone, véase tabla 2, un borrador sobre la síntesis de las inferencias y garantías que funcionen como guía para la elaboración de exámenes de ingreso a la universidad (Sánchez-Mendiola & Delgado-Maldonado, 2017), comprendiendo su complejidad y que se debe estimar por área de trabajo (español, matemáticas, lógica u otros). En este sentido, esta propuesta tiene la intención de ir mejorando para próximas investigaciones con relación a exámenes de ingreso a la universidad.

Tabla 2

Resumen de inferencias y garantías como modelo de argumento de validación de un examen de ingreso a la universidad

Inferencia	Garantía
Descripción del dominio	Observaciones de los conocimientos, destrezas y habilidades representativas a partir del uso objetivo del tema en la práctica universitaria.
Evaluación	Estimación apropiada de los puntajes para expresar rendimiento.
Generalización	Estimaciones estables del rendimiento examinado.
Explicación	Las puntuaciones se analizan según el comportamiento estadístico esperado.
Extrapolación	El constructo de la competencia definida es observable en el rendimiento futuro durante los estudios de educación superior.
Utilización	Los resultados de la prueba reflejan las habilidades previstas, uso a validar o afirmación a defender.

Fuente: Adaptación propia retomada de Chapelle, et al. (2010).

CONCLUSIÓN

Si bien se puede seguir discutiendo el término de validez se puede decir que en la actualidad los autores retoman la definición de los Estándares (AERA, APA & NCME, 2014). Asimismo, cuando hablamos de validez podemos referirnos tanto a investigaciones cualitativas como cuantitativas (Cohen, 2007; Kerlinger y Lee, 2001). Por otra parte, no es correcto hablar de tipos de validez si no de tipos de evidencia de validez, lo que cambia nuestra perspectiva al concepto de validez como algo unitario (AERA, APA & NCME, 2014).

En cuanto a la línea temporal del concepto podemos situarlo en cinco etapas, donde la sexta estaría sucediendo en estos momentos, estos son: el gestacional, de cristalización, de fragmentación, de reunificación y de deconstrucción. Siendo este último donde destacan los autores como Messick y Kane, el primero considerando a la validez de constructo y el segundo con su propio marco referencial a partir de sus cuatro tipos de inferencias para así desarrollar un sólo argumento de validez. Para ello debemos tomar en cuenta que la AERA, APA y el NCME (2014) consideran que la validez es más robusta cuando hay más evidencias. Esto hace que el argumento de validez de Kane sea propicio para el análisis y propuestas como la de Chapelle, et al., (2010). No obstante, algunas limitaciones de este análisis parten de no incluir un mapeo sistemático de la literatura o una revisión sistemática de la literatura que apoye el estado del conocimiento actual sobre el concepto de validez.

Por otra parte, el argumento de validez confirma su solidez a partir del IUA con ayuda del modelo de Toulmin (1958) y de las inferencias que se realicen; puntuación, generalización, extrapolación e implicaciones (Kane, 2001, 2013). En el caso de Chapelle, et al. (2010), se proponen otro tipo de inferencias, garantías y suposiciones que ayuden al proceso de validación del TOEFL; descripción del dominio, evaluación, generalización, explicación, extrapolación y utilización. De estos se elaboraron

referencias y definiciones para la prueba TOEFL en el ingreso a la universidad del idioma inglés. No obstante, como estaban descritos no son generalizables por lo que se propone un borrador de un modelo de argumento de validez donde se incluyen las inferencias y garantías para la elaboración de exámenes de ingreso a la universidad basados en la propuesta de Chapelle, et al. (2010).

Para cerrar, esta publicación es el inicio de una serie de publicaciones relacionadas a la validez de exámenes de ingreso a la universidad por lo que se seguirá profundizando en un modelo idóneo de validez para ello. Es decir, tenemos que seguir considerando estudiar qué ha pasado después de la última actualización de los Estándares de la AERA, APA y la NCME (2014), además de estructurar el método del modelo del enfoque basado en argumentos propuesto por Kane y después por Chapelle, su discípula.

REFERENCIAS

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). Estándares para pruebas educativas y psicológicas (M. Lieve, Trans.). American Educational Research Association.

Carrillo, B.; Sánchez, M., & Leenen, I. (2020). El concepto moderno de validez y su uso en educación médica. Investigación en Educación Médica, 98-106. https://doi.org/10.22201/facmed.20075057e.2020.33.19216

Chapelle, C. (2021). Argument-Based Validation in Testing and Assessment. SAGE.

Chappelle, C., Enright, M., & Jamieson, J. (2008). Building a Validity Argument for the Test of English as a Foreign Language. Routledge.

Chappelle, C., Enright, M., & Jamieson, J. (2010). Does an Argument-Based Approach to validity make a difference? Educational Measurement: Issues and Practice, 29(1), 3-13.

Cohen, L., Manion, L. & Morrison, K. (2007). Research methods in education. Routledge.

Cook, D., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. Medical Education, 49, 560-575. doi: 10.1111/medu.12678

Cronbach, L.J. (1951). «Coeficiente alfa y la estructura interna de los test (Coefficient alpha and the internal structure of tests)». Psychometrika (en inglés) 16 (3): 297-334. ISSN 0033-3123. Consultado el 14 de abril de 2012.

Delgado-Rico, E.; Carretero-Dios, H., & Ruch, W. (2012). Content validity evidences in test development: An applied perspective. International Journal of Clinical and Health Psychology, 12(3), 449-459. https://www.redalyc.org/pdf/337/33723713006.pdf

García-Medina, A.; Martínez-Rizo, F.; Cordero-Arroyo, G., & Caso-Niebla, J. (2017). Evolución del concepto de validez en la medición educativa. https://www.researchgate.net/publication/325346472_Evolucion_del_concepto_de_validez_en_la_medicion_educativa

Kane, M. (2006). Current Concerns in Validity Theory. Journal of Educational Measurement, 38(4), 319-342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2011). Validating score interpretations and uses. Language Testing, 29(1), 3–17. doi:10.1177/0265532211417210

Kane, M. (2013). Validating the interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38(4), 319-342. https://www.jstor.org/stable/1435453

Kerlinger, F. y Lee, H. (2001). Investigación del comportamiento: métodos de investigación en ciencias sociales. McGraw Hill.

López García, Y. (2019). A validity study of the Evaluation Infantil Temprana (EIT). [Tesis doctoral]. The University of New Brunswick.

Messick S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher (18), 2, 5-11.

Newton, P., & Shaw, S. (2014). Validity in educational & psychological assessment. SAGE.

Pedrosa, I., Suárez-Álvarez, J., & García-Cueto, E. (2013). Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. Acción Psicológica, 10(2), 3-18. https://dx.doi.org/10.5944/ap.10.2.11820

Popham, W. J. (2008). Transformative Assessment. Alexandria, VA: ASCD.

Rodríguez, B. (2004). El modelo argumentativo de Toulmin en la escritura de artículos de investigación educativa. Revista Digital Universitaria, 5(1), 2-18. https://www.revista.unam.mx/vol.5/num1/art2/ene_art2.pdf

Sánchez-Mendiola, M., & Delgado-Maldonado, L. (2017). Exámenes de alto impacto: implicaciones educativas. Investigación En Educación Médica, 6(21), 52–62. doi:10.1016/j.riem.2016.12.001

Taylor, C. (2013). Validity and Validation. Oxford.

Toulmin, S. (1958). The uses of argument. Cambridge University Press.

Watson, P. (2002). Introducción: la evolución de las leyes del pensamiento, en Historia intelectual del siglo XX, 11-15. Crítica.